

# **Information theoretical approaches for the identification of potentially cooperating transcription factors**

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades  
*"Doctor rerum naturalium"*  
der Georg-August-Universität Göttingen

im Promotionsprogramm Computer Science (PCS)  
der Georg-August University School of Science (GAUSS)

vorgelegt von

Cornelia Meckbach  
aus Kassel

Göttingen  
2019

### Betreuungsausschuss

Prof. Dr. Edgar Wingender,  
Institut für Bioinformatik, Universitätsmedizin Göttingen.

Prof. Dr. Stephan Waack,  
Institut für Informatik, Georg-August Universität Göttingen.

Dr. Mehmet Gültas,  
Department für Nutztierwissenschaften, Georg-August Universität Göttingen.

### Mitglieder der Prüfungskommission

Referent: Prof. Dr. Edgar Wingender,  
Universitätsmedizin Göttingen.

Korreferent: Prof. Dr. Stephan Waack,  
Georg-August Universität Göttingen.

Korreferent: Prof. Dr. Ralf Hofestädt,  
Universität Bielefeld

### Weitere Mitglieder der Prüfungskommission

Prof. Dr. Tim Reißbarth,  
Institut für medizinische Bioinformatik, Universitätsmedizin Göttingen.

Prof. Dr. Burkhard Morgenstern,  
Institut für Mikrobiologie und Genetik, Abteilung für Bioinformatik, Georg-August-Universität Göttingen.

Prof. Dr. Carsten Damm,  
Institut für Informatik, Georg-August-Universität Göttingen.

Prof. Dr. Rolf Daniel,  
Institut für Mikrobiologie und Genetik, Genomic and Applied Microbiology and Göttingen  
Genomics Laboratory, Georg-August-Universität Göttingen.

Tag der mündlichen Prüfung: 21.06.2019

## Abstract

Transcription factors (TFs) are a special class of proteins that usually bind regulatory DNA regions such as promoters and enhancers in order to control the expression of their target genes. Today, it is well known that in higher organisms, the combinatorial interplay between TFs is crucial for a flexible and precise gene regulation. Thereby, the cooperation between TFs is highly diverse and can take place between TFs that are bound to the same DNA region, referring to intra-regional TF cooperations as well as between TFs that are bound to different DNA regions (i.e. enhancer and promoter regions), referring to inter-regional TF cooperations. The computational identification of these TF cooperations is still a challenging problem in bioinformatics and can be addressed by using predicted transcription factor binding sites (TFBSs) as basis of the analysis. In this thesis, I present two information theoretical approaches for the identification of cooperating TFs based on their TFBS distributions in regulatory DNA regions.

My first approach identifies potentially intra-regional cooperating TFs based on the co-occurrence of their binding sites. Thereby, I adapted the pointwise mutual information from the field of linguistics to the field of bioinformatics by using it for the identification of co-occurring TFBSs. For this, I consider the genome as a document, the sequences under study as sentences and the predicted TFBSs as words in these sentences. I successfully applied this approach to a simulation data set, biological data sets and performed a comparison study with existing methods. Although the results reveal that my approach properly identifies known and novel TF cooperations, the differentiation between sequence-set specific pairs and common/general important ones is missing. Addressing this point, I extended my method and created background sequence-sets to estimate the background co-occurrence of each TFBS pair, incorporated it in the calculation and classified the significant pairs as sequence-set specific or common ones. Applying this extended version to several gene sets, the overlap between the sequence-set specific pairs is considerably decreased in comparison to the original version.

In order to complement my first method, I established a second approach for the determination of inter-regional TF associations that might be involved in the interaction process between promoter and enhancer regions. This approach is based on the sequences of known promoter-enhancer interactions and estimates the association between TFBS distributions of different DNA regions based on multivariate mutual information (MMI). Thereby, I created background sequence sets by preserving the (olig-) nucleotide composition and directly incorporated them in the MMI computation as a third random variable. Considering this approach, I compared the performance of four different mutual information quantities. Finally, I demonstrated the performance of this approach by successfully applying it to simulation and biological data sets and by comparing it with an existing method.





## Zusammenfassung

Transkriptionsfaktoren (TFs) sind eine spezielle Gruppe von Proteinen, die an regulatorische DNA Regionen wie Promotoren oder Enhancer binden, um die Expression ihrer Zielgene zu kontrollieren. Heutzutage ist hinlänglich bekannt, dass in höher entwickelten Organismen das kombinatorische Zusammenspiel von TFs unerlässlich für eine flexible und präzise Genregulation ist. Dabei ist die Kooperation von TFs sehr divers und kann zwischen TFs stattfinden, die an die gleiche DNA-Region gebunden sind, im Folgenden intraregionale TF Kooperationen genannt, sowie zwischen TFs, die an unterschiedliche DNA-Regionen gebunden sind (z.B. Enhancer- und Promotorregionen), im Folgenden interregionale TF-Kooperationen genannt. Die computergestützte Identifizierung dieser TF-Kooperationen ist nach wie vor ein herausforderndes Problem in der Bioinformatik und kann dadurch adressiert werden, dass vorhergesagte Transkriptionsfaktorbindestellen (TFBSs) im Hinblick auf ihr gemeinsames Auftreten analysiert. In dieser Arbeit präsentiere ich zwei informationstheoretische Verfahren für die Identifikation von kooperierenden TFs basierend auf deren TFBSs-Verteilungen in regulatorischen DNA-Regionen.

Mein erstes Verfahren identifiziert potenzielle intraregionale TF-Kooperationen basierend auf dem gemeinsamen Vorkommen ihrer Bindestellen. Dabei habe ich die *pointwise mutual information* aus der Linguistik für die Bioinformatik angepasst, um gemeinsam vorkommende TFBSs vorherzusagen. Hierfür betrachte ich das Genom als ein Dokument, die zu analysierenden Sequenzen als Sätze und die vorhergesagten TFBSs als Wörter in diesen Sätzen. Ich habe das Verfahren erfolgreich auf einen simulierten Datensatz und auf biologische Datensätze angewendet und eine Vergleichsstudie mit bereits existierenden Methoden durchgeführt. Obwohl die Ergebnisse zeigen, dass meine Methode bereits bekannte und neue TF-Kooperationen erfolgreich identifiziert, fehlt die Unterscheidung zwischen solchen Paarungen, die für den jeweils untersuchten Sequenz-Set spezifisch sind, und solchen, die allgemein wichtig sind und daher stets in Erscheinung treten. Um diesen Punkt zu berücksichtigen, erweiterte ich die Methode und erzeugte Hintergrundsequenzsets um die Hintergrundkolokalisation für jede TFBS-Paarung abzuschätzen und dieses in meine Berechnung zu integrieren, um somit die signifikanten Paarungen als Sequenz-Set-spezifisch oder allgemein wichtig zu klassifizieren. Die Anwendung dieser erweiterten Methode auf unterschiedlichen Gensets zeigt, dass die Überlappung zwischen Sequenz-Set spezifischen Paarungen wesentlich geringer ist im Vergleich zu der originalen Methode.

Mit dem Ziel, die erste Methode zu komplementieren wurde ein zweites Verfahren entwickelt, dass interregionale TF Beziehungen ermitteln soll, welche möglicherweise in den Interaktionsprozess zwischen Enhancer- und Promotorregionen involviert sind. Dieses Verfahren basiert auf den Sequenzen von bekannten Promoter-Enhancerinteraktionen und schätzt die Assoziation zwischen TFBS Verteilungen unterschiedlicher DNA-Regionen

mittels der *multivariate mutual information* (MMI) ab. Dabei werden Hintergrundsequenzen erzeugt, bei denen die (Oligo-)Nukleotidzusammensetzung erhalten bleibt und die direkt als dritte Zufallsvariable in die MMI-Berechnung mit eingefügt werden. Für dieses Verfahren habe ich die Performance von vier unterschiedlichen MMI-Metriken miteinander verglichen. Abschließend demonstrierte ich die Leistung dieses Verfahrens, indem ich es erfolgreich auf simulierte sowie auf biologische Datensätze angewendet habe und mit einer bereits existierenden Methode verglichen habe.

## Acknowledgements

During my PhD study, I was accompanied by a lot of amazing people and I thank all of them for their contributions in many different ways.

First of all, I would like to thank Prof. Dr. Edgar Wingender who offered me the opportunity to do my PhD in his department. All the time, Prof. Wingender was open for questions, inspiring discussions and when ever I had new ideas he supported and motivated me to try them out. Thereby, he always provided a warm and welcoming atmosphere and opened my mind to see the bigger picture of science. Thank you!

I would also like to thank my second supervisor Prof. Dr. Stephan Waack for his support with the mathematical modeling, presenting to me the information theory from the mathematical point of view and investing a lot of time in nice and inspiring discussions where he showed me points I would not even have thought about.

Further, I would like to acknowledge all the members of my thesis committee: Prof. Dr. Tim Beißbarth, Prof. Dr. Carsten Damm, Prof. Dr. Rolf Daniel and Prof. Dr. Burkhard Morgenstern. They spend their valuable time on me, thanks a lot for that!

During the time of my master thesis and my PhD time I am extremely grateful to have Dr. Mehmet Gültas as my supervisor. Dr. Gültas put a lot effort in my PhD and our publications and supported me in every sense. He kept supervising me although he changed his position. Mehmet, it was my great pleasure to work with you!

I also thank all the people from the former and current *Institute of (medical) bioinformatics* for their help and support. Thereby, I am very grateful for Doris for her astounding help and support, Sebastian for the nice work and non-work discussions, Torsten for technical support and last but not least for Rayan who supplied me nearly every lunch time with a new *rice and chicken* creation and self-made sweets. Rayan, I will miss our acro yoga lunch breaks!!!

I further give my thanks to Maren, Becky and Simeon for proofreading this thesis, to all my climbing-friends, the Stoppelhopper-team and all the new friends I made in Göttingen for the unforgettable time.

Finally, I would like to thank my parents. Without questioning they always gave their best to support me during my studies and my daily live. I dedicate this thesis to them.



# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Structure of the thesis . . . . .	3
1.2. Impact . . . . .	3
<b>2. Biological background</b>	<b>7</b>
2.1. The molecular mechanisms of gene expression . . . . .	7
2.1.1. DNA stores the genetic information . . . . .	7
2.1.2. Gene expression: decoding of genetic information . . . . .	9
2.1.3. Regulation of gene expression . . . . .	9
2.1.4. Transcription and its regulation . . . . .	10
2.2. Experimental methods . . . . .	15
2.2.1. Determination of TFBSs . . . . .	16
2.2.2. Determination of promoter-enhancer interactions . . . . .	16
2.3. Bioinformatic resources . . . . .	18
2.3.1. Bioinformatic data bases . . . . .	18
2.3.2. Bioinformatic tools . . . . .	27
<b>3. Theoretical background</b>	<b>31</b>
3.1. Information theory . . . . .	31
3.1.1. Entropy . . . . .	31
3.1.2. Mutual Information . . . . .	33
3.1.3. Multivariate mutual information . . . . .	36
3.1.4. Pointwise mutual information . . . . .	38
<b>4. Information theoretical approaches for the analysis of cooperating TFs</b>	<b>43</b>
4.1. Identification of intra-regional cooperating TFs using pointwise mutual information . . . . .	43
4.1.1. Cooperating TFs . . . . .	44
4.1.2. Sequence-set specific cooperating TFs . . . . .	50
4.2. Identification of inter-regional associated TFs using multivariate mutual information . . . . .	55

---

<b>5. Results</b>	<b>65</b>
5.1. Identification of intra-regional cooperating TFs using pointwise mutual information . . . . .	65
5.1.1. Cooperating TFs . . . . .	65
5.1.2. Sequence-set specific cooperating TFs . . . . .	76
5.2. Identification of inter-regional associated TFs using multivariate mutual information . . . . .	87
5.3. Identification of inter- and intra-regional cooperating TFs in the context of inflammatory response in lung tissue . . . . .	107
<b>6. Discussion</b>	<b>113</b>
6.1. Pointwise mutual information in the context of intra-regional cooperating TF identification . . . . .	113
6.2. Multivariate mutual information in the context of inter-regional cooperating TFs . . . . .	117
6.3. Complementarity of PMI and MMI in a biological application . . . . .	120
6.4. Impact of combinatorics in transcription regulation . . . . .	121
<b>7. Conclusion</b>	<b>123</b>
7.1. Summary . . . . .	123
7.2. Outlook . . . . .	124
<b>Bibliography</b>	<b>127</b>
<b>A. Appendix</b>	<b>144</b>
A.1. PC-TraFF: identification of potentially collaborating TFs using pointwise mutual information . . . . .	144
A.2. Removing background Co-occurrences of TFBSs greatly improves the prediction of specific TF cooperations . . . . .	166
A.3. Computational detection of stage-specific TF clusters during heart development . . . . .	178
A.4. A novel sequence-based feature for the identification of DNA-binding sites in proteins using Jensen-Shannon Divergence . . . . .	196

## List of Figures

2.1. Structure of DNA . . . . .	8
2.2. Process of gene expression with the levels of regulatory mechanisms . . . . .	10
2.3. Preinitiation complex of RNA polymerase II . . . . .	11
2.4. Polymerase II core promoter . . . . .	12
2.5. Mechanisms determining promoter-enhancer interactions . . . . .	13
2.6. Physical cooperation strategies of transcription factors . . . . .	14
2.7. Modular composition of transcription factors . . . . .	14
2.8. Strategies of repressing transcription factors . . . . .	15
2.9. Nuclease protection footprinting . . . . .	17
2.10. The basic structure of TRANSFAC database . . . . .	19
2.11. Screenshot of the ENCODE search interface . . . . .	22
2.12. Screenshot of the organism selection menu of the UCSC Genome Browser . . . . .	23
2.13. Screenshot of the interface of the UCSC Table Browser . . . . .	24
2.14. Screenshot of the result page for SP1 protein interactions in human . . . . .	25
2.15. Screenshot of STRING result page of SP1 protein in human . . . . .	26
2.16. Example of a MATCH <sup>TM</sup> output . . . . .	28
3.1. Relation between entropy and mutual information . . . . .	34
3.2. Relation between entropy, mutual information and multivariate mutual information . . . . .	36
3.3. Information theory measures for three random variables. . . . .	39
3.4. Maximized $\mathbb{P}^{\text{PMI}}$ in dependence of occurrence probability $p(x)$ of $x$ . . . . .	41
4.1. Construction of TFBS-sequence matrix . . . . .	45
4.2. Different scenarios for overlapping TFBSs . . . . .	46
4.3. Filter to avoid overlaps. . . . .	47
4.4. TFBS pair construction . . . . .	49
4.5. Workflow of the extension approach for the determination of sequence set specific TFBS pairs . . . . .	53
4.6. Identification of associated TFs between enhancer and promoter sequences using mutual information $\mathbb{I}$ . . . . .	56
4.7. Determination of TFBS-sequence count matrices . . . . .	57
4.8. Conversion of TFBS-sequence matrix to interval-sequence matrix. . . . .	61

---

5.1. Cooperation network of PC-TraFF significant TFBS pairs of whole genome analysis . . . . .	72
5.2. Cooperation network of PC-TraFF significant TFBS pairs of breast cancer gene set analysis . . . . .	77
5.3. Number of specific TFBS pairs in dependence on different $\alpha$ -values . . . . .	79
5.4. Logoplot alignment for the TFBSs involved in the four top ranking pairs . . . . .	80
5.5. Number of unique and overlapping significant TFBS pairs for the different breast cancer subtypes . . . . .	81
5.6. Number of sequence-set specific TFBS pairs for the five breast cancer subtypes . . . . .	84
5.7. Cooperation network of Luminal A significant TFBS pairs according to the original method . . . . .	85
5.8. Cooperation network of Basal-like significant TFBS pairs according to the original method . . . . .	86
5.9. Example dataset . . . . .	88
5.10. PEI sub-network . . . . .	98
5.11. Length distribution of enhancer and promoter sequences for K562 cell line . . . . .	99
5.12. Number of unique and overlapping single TFBSs participating in significant pairs of the different cell lines . . . . .	100
5.13. Number of unique and overlapping significant TFBS pairs of the different cell lines . . . . .	101
5.14. Degree distribution of nodes of the K562 cooperation network . . . . .	102
5.15. TFBS association network between enhancer and promoter regions for cell line K562 . . . . .	104
5.16. Joint network of inter- and intra-regional cooperating TFs . . . . .	109



## List of Tables

2.1.	Latest public statistics of TRANSFAC <sup>®</sup> . . . . .	21
2.2.	BioGRID statistics of January 2019 . . . . .	22
2.3.	Statistics of TRANSCompel <sup>®</sup> . . . . .	24
3.1.	Difference between pointwise mutual information $\mathbb{P}MII$ and mutual information $MII$ . . . . .	40
5.1.	Total number of predicted TFBS pairs for the genome wide and the breast cancer analysis . . . . .	67
5.2.	Pairwise comparison of the different approaches. . . . .	68
5.3.	Performance comparison of the different approaches . . . . .	69
5.4.	Combination of the different approaches . . . . .	70
5.5.	Significant TFBS pairs found by the method in genome-wide promoter analysis of human RefSeq genes . . . . .	71
5.6.	Exemplary comparison between the TFBSs contained in the left and the right cluster . . . . .	72
5.7.	The hubs and their top three collaboration partners . . . . .	74
5.8.	The hubs and their top three collaboration partners . . . . .	75
5.9.	Number of promoter sequences of breast cancer subtype-associated RefSeq genes and corresponding significant pairs found by my approach . . . . .	77
5.10.	Total number of sequence-set specific TFBS pairs for the simulation dataset for different $\alpha$ -values . . . . .	78
5.11.	Six significant TFBS pairs determined as significant by the original approach for all breast cancer subtypes. . . . .	82
5.12.	Pairs that were identified as significant . . . . .	83
5.13.	Results of the synthetic generated count matrices . . . . .	89
5.14.	Inserted associated TFBSs in enhancer and promoter sequences with the representing logoplots . . . . .	90
5.15.	Visualization of the different states of the “association strength” variable . . . . .	91
5.16.	Numbers of inserted TFBS instances for each artificially inserted associated TFBS pairing . . . . .	92
5.17.	Number of significant pairs identified by $MIII$ for the simulation dataset of each condition . . . . .	93
5.18.	Results for the simulation dataset . . . . .	94

---

5.19. Results of MotifHyades in comparison to my approach using the <i>MMII</i> . . .	96
5.20. Number of enhancers, promoters and PEIs for the different cell lines . . . .	97
5.21. Average length of promoter and enhancer sequences for each cell line . . .	97
5.22. Summary of the identified inter-regional TFBS pairs using <i>MMII</i> for the different cell lines . . . . .	98
5.23. Highly associated TFBSs of the identified inter-regional TFBS pairs for the different cell lines . . . . .	102
5.24. Top ten associated TFBS pairs for cell line K562 . . . . .	105
5.25. Summary of the cooperation networks based on the intra- and inter-regional analyses . . . . .	107
5.26. Hub nodes for the inter-and intra regional cooperating TFBS network . . .	108
5.27. TFBSs identified in the analysis for inter-regional and intra-regional TF co- operations . . . . .	110

# Acronyms

<b>APC</b>	Average product correction
<b>CMI</b>	Conditional mutual information
<b>CML</b>	Chronic myelogenous leukemia
<b>COPS</b>	Co-occurrence pattern search (A tool for the detection of co-occurring transcription factors)
<b>CPModule</b>	A tool for the detection of cis-regulatory modules.
<b>CSS</b>	Core similarity score
<b>DNA</b>	Deoxyribonucleic acid
<b>DTC</b>	Dual total correlation
<b>GTF</b>	General transcription factor
<b>I</b>	Mutual information
<b>JMI</b>	Joint mutual information
<b>MCC</b>	Matthews correlation coefficient
<b>MMI</b>	Multivariate mutual information
<b>mRNA</b>	Messenger ribonucleic acid
<b>miRNA</b>	Micro RNA
<b>MSS</b>	Matrix similarity score
<b>PEI</b>	Promoter-enhancer interaction
<b>PMI</b>	Pointwise mutual information
<b>PWM</b>	Position weight matrix
<b>RNA</b>	Ribonucleic acid
<b>snRNA</b>	Small nuclear RNA
<b>tRNA</b>	Transport RNA
<b>TF</b>	Transcription factor
<b>TFBS</b>	Transcription factor binding site
<b>TSS</b>	Transcription start site



# 1. Introduction

A flexible and specific gene regulation enabling the control of different genetic programs such as organogenesis, immune response and adaptation to environmental conditions is crucial for the survival, development and general fitness of an organism. The major control level of gene expression is transcription regulation which underlies the interplay between a multitude of regulatory DNA regions such as promoters and enhancers. While promoters are mostly directly upstream of the transcription start site (TSS) of a gene, enhancers can be millions of base pairs away from their target genes but come in close proximity to the promoter by the formation of chromatin loops, which are stabilized by interactions between proteins positioned at the one and the other of these regions. Thereby, the pairing between an enhancer and a promoter has been detected as highly tissue specific and is therefore of major importance for tissue development [1]. The regulatory DNA regions are occupied by transcription factors, a special class of proteins that specifically bind to defined DNA motifs that are referred to as transcription factor binding sites (TFBSs). Since in higher organisms, the number of genes strongly exceeds the number of transcription factors, their combinatorial binding and physical as well as functional interactions are of major importance for a proper gene regulation. Therefore, TFs tend to form dimers (as homo- or heteromers) or higher order complexes in order to synergistically or antagonistically influence the transcription of their target gene. Further, the combination of bound TFs and the interplay between the underlying factors is essential to establish the pairing between enhancer and promoter regions. Thereby, intra-regional cooperating TFs are referred to TFs that are bound to the same DNA region whereas inter-regional cooperating TFs are linked to associated TFs between enhancer and their related promoter regions.

The knowledge about interacting TFs is crucial in the general understanding of the molecular mechanisms underlying gene regulation and can further be used for the identification of important key players in these regulatory mechanisms. The computational identification of interacting TFs is still a challenging problem in bioinformatics. Several existing approaches identify cooperating TFs based on their binding site distribution in the regulatory sequences under study. Thereby, most of these methods [2, 3, 4, 5, 6, 7, 8, 9, 10, 11] focus on intra-regional cooperations of transcription factors and require user provided negative and positive control sets as well as previous knowledge about transcription factor interactions. Some other methods are restricted to simple organisms or small input sequence sets and are thereby limited in their general usage [3, 6, 8, 12]. For example, Girgis et al. developed a tool for the identification of enriched motif pairs using a Bayesian classifier in a given set of sequences in comparison to a user provided control set [13]. Another approach has been

developed by Sun et al. [7] for the detection of unstructured cis-regulatory modules based on constrained programming for itemset mining framework that uses the whole genome as background sequence set. In 2013 Deyneko et al. [4] developed a method for the identification of composite elements that are stored in the TRANSCompel<sup>®</sup> [14] database. Thereby, the algorithm scans the input sequences and outputs the predicted locations of composite elements. However, the algorithm is not able to identify new composite elements that have not been experimentally verified yet.

To overcome the obstacles of the existing methods (e.g. user provided or arbitrary background set, restriction to known cooperations of TFs or data size in general), I propose a new method for the identification of cooperating transcription factors based on the co-occurrence of their underlying TFBSs. Inspired by the field of linguistics, where the pointwise mutual information (PMI) is a powerful tool for the identification of word associations, I adopted the PMI to the field of bioinformatics. Thereby, I consider the genome as a book, the regulatory sequences under study as sentences and TFBSs as words in these sentences. The results show that the application of pointwise mutual information in bioinformatics successfully determines the inserted pair in a simulation dataset. In application to biological sequences, it is able to identify known TF cooperations as well as new potential TF cooperations which could provide new targets for future laboratory work. Although the predicted pairs appear to be important for the regulation of the underlying gene set, the overlap of significant pairs between different input sets is comparatively huge. This indicates that the predicted pairs can be divided into two groups: sequence-set specific pairs and common important ones that stem from generally used regulatory programs in many cells and tissues. In order to separate the predictions into sequence-set specific and common ones, I extended my approach by creating background sequence sets that maintain the (oligo-) nucleotide composition of the input sequences, estimating the background co-occurrence for each TFBS pair and subtracting this background from the original pointwise mutual information value. A closer look at the predictions reveals that the overlap of sequence-set specific TFBS pairs among different input sets decreases in comparison to the original approach, pointing out the success of the extended approach.

Up to date, only a few computational methods exist for the identification of coupled transcription factors that are involved in the pairing process between enhancer and promoter sequences. A recent approach in this field is MotifHyades [15], proposed by Wong in 2017, for the identification of coupled DNA motif pairs in enhancer and promoter sequences based on expectation maximization methodology. This probabilistic method performs well for the identification of over-represented pairs, however, the algorithm is not able to identify associated motif pairs that only occur in a minority of the promoter-enhancer pairings.

In this thesis, I present a new approach for the identification of associated transcription factors in enhancer and promoter sequences based on multivariate mutual information. For this, a background sequence set was created by maintaining the general sequence (oligo-)

nucleotide composition and afterwards, the distribution of TFBSs in both, input and background enhancer and promoter sequences was calculated, respectively. Later, the pairwise association between a TFBS of promoter sequences and a TFBS of enhancer sequences was calculated by mutual information where further, the background information was incorporated as third random variable in the analysis. In order to find the best mutual information metric for my purposes, I compared and evaluated several quantities (i.e. dual total correlation, conditional mutual information, multivariate mutual information and pairwise mutual information of joint distributions) that consider three random variables and conclude that the multivariate mutual information is the best choice for the identification of associated transcription factors, since it identifies strong and weak associated TFBS pairs in the underlying promoter-enhancer interactions.

## 1.1. Structure of the thesis

The organization of the thesis is as follows. In Chapter 2, I introduce the most relevant biological facts about gene regulation by focusing of transcriptional regulation and transcription factors. I further introduce some experimental and bioinformatical methods that are related to the data in the thesis and give an overview about the bioinformatical resources and data bases used in this thesis. In Chapter 3, I give a brief overview about information theory and entropy and focus afterwards on different mutual information quantities. Followed by this foundation chapters, I introduce the information theoretical approaches established in this thesis in Chapter 4. Thereby, I first present the method for the identification of potentially intra-regional cooperating TFs based on pointwise mutual information in Section 4.1 in combination with the extended version of this approach for the identification of sequence-set specific TF cooperations. In the following Section 4.2, I describe the multivariate mutual information based method for the identification of associated TFBSs between promoter and their related enhancer regions based on their underlying TFBS distributions. Afterwards, I applied both methods to simulation and real biological data sets and present the results in combination with comparative studies to existing methods, in Chapter 5. These results as well as the application of the information theoretic methodology is discussed in Chapter 6 and finally, I complete the thesis in Chapter 7 by summarizing the thesis and give an outlook for future work.

## 1.2. Impact

### Journal articles:

We have published the pointwise mutual information based method for the identification of intra-regional TF cooperations as well as its extension for sequence-set specific TF cooperations in the following articles:

- [1] **Meckbach, C**, Tacke, R, Hua, X, Waack, S, Wingender, E, Gültas, M (2015). *PC-TraFF: identification of potentially collaborating transcription factors using point-wise mutual information*. BMC Bioinformatics, 16:400.
- [2] **Meckbach, C**, Wingender, E, Gültas, M (2018). *Removing Background Co-occurrences of Transcription Factor Binding Sites Greatly Improves the Prediction of Specific Transcription Factor Cooperations*. Front Genet, 9:189.
- [3] Steuernagel, L\*, **Meckbach, C\***, Heinrich, F, Zeidler, S, Schmitt, A, Gültas, M (2019). *Computational identification of tissue-specific transcription factor cooperation in ten cattle tissues*. PLoS ONE, accepted 29.4.2019 and currently print (\*These authors contributed equally to this work.).

Further, the author contributed to the following publications that are related to the topic of the thesis:

- [4] Zeidler, S, **Meckbach, C**, Tacke, R, Raad, FS, Roa, A, Uchida, S, Zimmermann, WH, Wingender, E, Gültas, M (2016). *Computational Detection of Stage-Specific Transcription Factor Clusters during Heart Development*. Front Genet, 7:33.
- [5] Dang, T.K.L., **Meckbach, C.**, Tacke, R., Waack, S. and Gültas, M (2016).: *A novel sequence-based feature for the identification of DNA-binding sites in proteins using Jensen–Shannon Divergence*. Entropy 18:379.

### Conferences, Workshops, Meetings and Student's thesis

The author represents topics of this thesis on the following workshops and conferences:

- European Conference on Computational Biology (ECCB 2016, September The Hague): Poster presentation
- German Conference on Bioinformatics (GCB 2016, Berlin): Poster presentation
- Bioinformatic poster day (Göttingen 2017): Poster presentation
- Workshop on Bioinformatics of Gene Regulation (Göttingen 2018): Poster presentation and talk

In collaboration with Mehmet Gültas and Edgar Wingender the author supervised the following student works:

- Felix Heinrich: *PC-TraFF Matchscores: Miteinbeziehung von TF-Bindestellenqualität bei der Bestimmung von interagierenden TFs sowie die Identifizierung ihrer bevorzugten Bindestellendistanzen*. Bachelor Thesis, 2016
- Lena Steins: *Analyzing transcription factor interactions in the embryonic development of human cardiomyocytes using PC-TraFF*. Master Thesis, 2016-2017
- Selina Klees: *Analysis of Promoter-Enhancer Interactions by comparing the Transcription Factor Binding Site Composition*. Project Work, 2017



- Lukas Steuernagel: *Modellierung des Informationsgehalts von eukaryotischen und prokaryotischen Promotoren anhand von vorhergesagten Transkriptionsfaktorbindstellen und den dahinter stehenden Datenbankinformationen*. Project Work, 2017

In collaboration with Mehmet Gültas and Felix Heinrich, we further provide a web server for the identification of intra-regional cooperating TFs based on the approach of Section 4.1 that is available via <http://pctraffpro.bioinf.med.uni-goettingen.de/>.



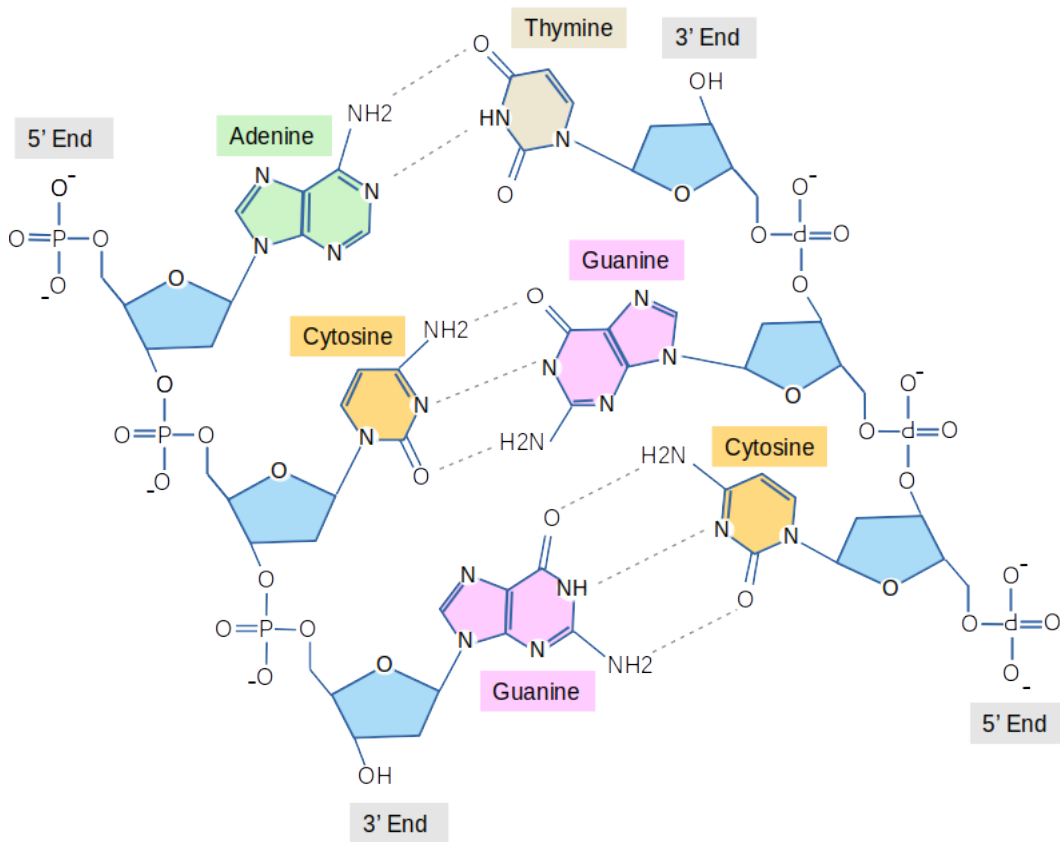
## 2. Biological background

In this chapter, I give an overview about the molecular processes and components in the cell that are required to fully understand the motivation and concepts of this thesis. Thereby, I will first give an overview about gene expression and regulation in general and more insights into transcription itself and the regulation of transcription governed by transcription factors. For a more detailed presentation of the biological parts I kindly refer to text books like [16, 17] and especially for transcription factors to [18]. In the last part of the chapter, I give an overview about the experimental methods, bioinformatic databases and tools that are required in this thesis for evaluation as well as the for the data generation as pre-processing work.

### 2.1. The molecular mechanisms of gene expression

#### 2.1.1. DNA stores the genetic information

Since 1940 [17] it is known that the deoxyribonucleic acid (DNA) is the cellular component that captures the genetic information of an organism. In 1953 the three dimensional structure of DNA was discovered under the direction of James Watson and Francis Crick. They found out that the DNA in general consists of two anti-parallel nucleotide chains that are twisted around each other forming a double helical structure. A DNA nucleotide consists of a sugar molecule (deoxyribose), a phosphate group and one of the four bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Building up the linear polymer, the sugar molecules are linked by the phosphate groups and form the uniform backbone of the helical structure, while the bases point inside the helix and are paired to the facing base. These base pairings (bp) are structurally determined by hydrogen bonds in a way that adenine pairs with thymine and guanine pairs with cytosine as illustrated in Figure 2.1. The phosphodiester bond between the nucleotides results in a defined orientation of the nucleotide chain defined by the phosphate end (linked to the 5' carbon of the deoxyribose) and the sugar end (defined by the free OH-group of the 3' carbon of deoxyribose). In literature, a DNA sequence is in general oriented from the 5' end to the 3' end leading to the terms *upstream* (towards the 5' end) and *downstream* (towards the 3' end).



**Figure 2.1.: Structure of DNA.** The DNA consists of two nucleotide strands that are anti-parallel orientated. Each nucleotide consists of a sugar molecule (blue), one of the four bases (cytosine, guanine, adenine and thymine) and a phosphat group that enables the linear polymerization of the nucleotides. The two nucleotide strands in turn are connected by hydrogen bonds formed by the pairings between guanine and cytosine (three hydrogen bonds) or adenine and thymine (two hydrogen bonds).

In the cell, a DNA molecule is associated with a multitude of proteins forming a molecular complex that is termed chromatin, while the chromatin of one long DNA molecule, in particular in its compact form during metaphase, is a chromosome. This complex formation of DNA and proteins compacts the DNA that it fits inside the cell. In addition, packing the DNA in chromosomes increases the stability of the DNA molecule and the associated proteins can influence the accessibility of the DNA molecule and thereby influence gene expression [16, Page 135].

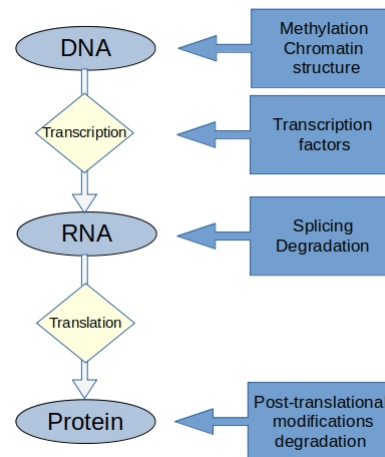
### 2.1.2. Gene expression: decoding of genetic information

A DNA region that codes for a functional molecule is termed gene and is a major constituent of holding the genetic information of an organism. The process leading to the decoding of the information, in order to form functional molecules, is termed gene expression and can be separated into two parts: transcription and translation [17]. During transcription, the gene sequence is transcribed into a ribonucleic acid (RNA) sequence and afterwards, translated into an amino acid sequence. RNA differs from DNA in some major ways: the sugar molecule of RNA is ribose and the base thymine is replaced by uracil. However, the major difference between the double stranded DNA and RNA is that in the cell, RNA occurs as a single stranded molecule that forms coiled and helical structures with itself [16]. RNA can fulfill regulatory and catalytic functions (miRNA, snRNA, rRNA, tRNA) or serve as a template for the synthesis of proteins (mRNA). In eukaryotes the product of transcription is a precursor of the final RNA, which in turn is generated by further processing in RNA during which the ends of the RNA are modified and intron parts are spliced out of the RNA sequence. The final RNA product is transported from the nucleus into the cytoplasm where the mRNA is translated into a polypeptide, a linear sequences of amino acids that form the main constituents of proteins [17].

### 2.1.3. Regulation of gene expression

The expression of some genes is continuously required while the products of some other genes are only needed under certain conditions (e.g. tissue development, environmental changes, etc.). In order to produce the right amount of gene products, the expression of a gene needs to be regulated. As shown in Figure 2.2, this regulation can take part on each step of gene expression.

The first major control level is the DNA structure that can be modified by methylation and the alteration of chromatin structure in a way that the accessibility of DNA for proteins is changed, which can result in the complete silencing of DNA regions. Further, the process of gene transcription is regulated by proteins termed transcription factors (TFs) that activate or repress the transcription of their target genes by usually binding to regulatory DNA sequences (see Section 2.1.4). The next control level of gene expression comprises the gene product itself where the speed of degradation of the transcribed precursor RNA determines the amount of the final gene product. In turn, the speed of degradation can be influenced by the length of the poly-A tail of RNA and the 5' RNA end capping process [16] and RNA splicing can lead to a multitude of different protein products of one RNA molecule. Finally, by post-translational modifications (e.g. phosphorylation) and, thus, the activity of the final protein as well as protein degradation can be regulated [18].

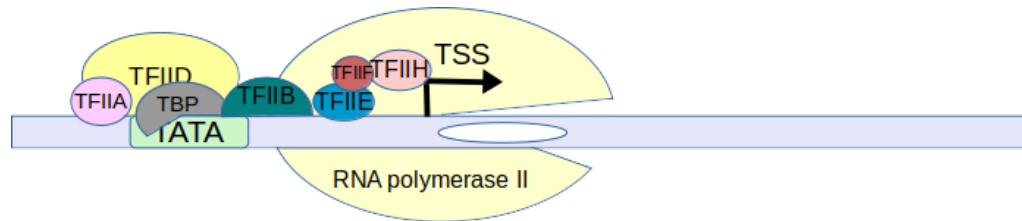


**Figure 2.2.: Process of gene expression with the levels of regulatory mechanisms.**

#### 2.1.4. Transcription and its regulation

**Transcription process** The transcriptional process in general is separated in three phases: initiation, elongation and termination. In the initiation phase, the RNA polymerase binds the DNA close to the transcription start side (TSS) in combination with general transcription factors (GTFs) supporting the formation of the pre-initiation complex that is depicted in Figure 2.3. This complex opens the DNA double helix and short RNA transcripts are synthesized by the RNA polymerase at the TSS [16]. After the first RNA transcript exceeds a length of about ten ribonucleotides the elongation phase starts that is simply the polymerisation of further ribonucleotides according to the DNA template by moving along the DNA strand [16]. The termination phase starts after the RNA polymerase passes the poly A signal sequence, the RNA strand is released, the RNA polymerase dissociated from the DNA and the transcription bubble is closed [16].

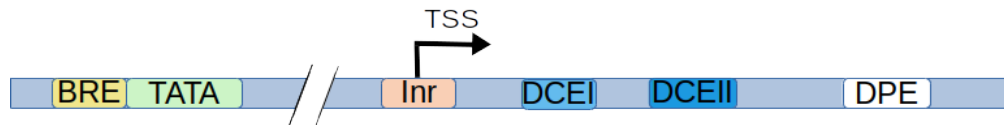
**Regulatory DNA regions** The transcription efficiency of a gene is influenced in *cis* by a couple of DNA regions such as promoters, enhancers, upstream activator sequences (UASs), insulators and boundary elements. In *cis* means that the regulatory element is on the same DNA molecule as the gene. A promoter is located immediately upstream of the transcription start side and can even reach within the coding region of the gene [18]. In eukaryotes, the totality of functional elements of the promoter (*cis* elements) that are sufficient to activate the transcription are referred to as core promoter and consists of 40-60 nucleotides in length [16]. The composition of these *cis* elements is specific and varies from gene to gene [19]. Common elements of eukaryotic RNA polymerase II core promoters are the TFIIB recognition elements (BRE), the TATA box, the initiator (Inr) as well as some downstream



**Figure 2.3.: Preinitiation complex of RNA polymerase II.** The binding of RNA polymerase II to the promoter is supported by general transcription factors denoted as TFII (transcription factors for RNA polymerase II) with classifications: TFIIA, TFIIB, TFIID, TFIIE, TFIIIF, TFIIH. The TATA box is recognized by the TATA-binding protein (TBP), a subunit of TFIID. (Modified from [16, Fig. 12-15])

promoter elements like the downstream promoter element (DPE), downstream core element (DCE) and the motif ten element (MTE) [16, Page 397] (see Figure 2.4). In general, a subset of these elements is sufficient to enable the binding of polymerase, general transcription factors and co-activators and thus, to enable the formation of the preinitiation complex [16, 19]. Besides the promoter, another important regulatory element is the enhancer, a cluster of regulatory sequences that is located hundreds or even millions of base pairs upstream or downstream from its target gene [16, 19, 20]. Enhancers form looping structures to physically interact with the promoter of their target gene irrespective of orientation [21], leading to transcription activation or the increase of the transcriptional level. The target genes of an enhancer can either be neighbouring genes but even the skipping of some genes is possible to reach their target genes [21]. Rarely, enhancer and target gene are located on different chromosomes [22]. The activity of enhancers is cell type specific or is affected by developmental or environmental constraints [19] indicating that the alterations of enhancer activities results in the change of gene expression patterns and consequently, incorrect alteration of enhancer activity are linked to many human diseases [23]. The enhancer activity itself can be identified by eRNAs, short non-coding RNAs that are bidirectionally transcribed from enhancer sequences if the enhancer elements are in close proximity to RNA polymerase II [19]. In addition, active enhancers can be identified by the proteins bound to them, i.e. they are often bound by the factor EP300 [21]. In the mammalian genome, there are around 23000 genes and about 1 million enhancers, indicating that several enhancers can act on the same target gene depending on the cell type or condition [19]. In turn, an enhancer can regulate several genes. The underlying mechanism of how an enhancer finds its target promoter is not fully understood yet. Following von Arensbergen et al. [21] mechanisms that might be involved in this selection process are: i) biochemical compatibility, ii) spatial architecture, iii) insulation and iv) chromatin environment. These mechanisms are illustrated in Figure 2.5. In detail, two regulatory sequences are biochemically compatible if both of them have the ability to be occupied by protein combinations that are able to

interact with each other. Obviously, the physical interactions between two sequences can only take place if the overall folding of the chromatin renders it possible. As mentioned above, another kind of *cis* regulatory DNA regions are insulator elements that can promote or block the interaction between an enhancer or a promoter by altering the 3D conformation of chromatin. These DNA regions are bound by specific DNA binding proteins where the most popular binding partner is the CTCF-binding factor (CTCF) [21].

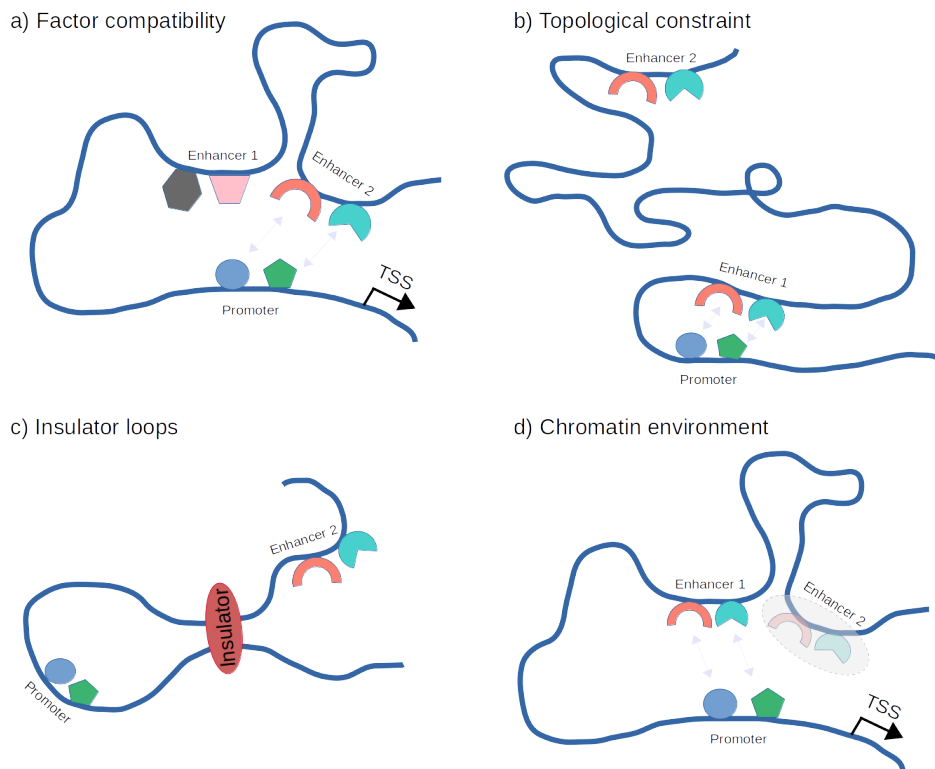


**Figure 2.4.: Polymerase II core promoter** with transcription start site (TSS) common regulatory elements: BRE (TFIIB recognition element), TATA (TATA Box), Inr (initiator element), DCE (downstream core element) and DPE (downstream promoter element). (Based on [16, Fig. 12-14])

**Transcription factors** In order to carry out their regulatory functions, the instructions encoded in the sequences of the *cis* regulatory elements are recognized by the selective binding of proteins to these regulatory sequence elements. These proteins belong to the overall class of transcription factors (TFs), regulatory proteins that are directly involved in the regulation process of a gene by usually binding to specific regulatory DNA sequences termed transcription factor binding sites (TFBSs) [25]. Fulfilling their regulatory functions, TFs can completely activate or repress transcription of a certain gene, or increase/decrease the level of its transcription. Thereby, TFs directly interact with the basal transcriptional machinery or alter chromatin structure by histone or DNA modifications. Regarding their molecular structure, TFs in general exhibit a modular composition (see Figure 2.7) and contain at least one of the following protein domains: i) a DNA binding domain, ii) an oligomerization domain, iii) a regulatory domain and iv) a trans-activation domain [26]. The DNA-binding domain recognizes specific DNA sequence patterns and enables the protein-DNA binding. DNA-binding domains of proteins can be computationally predicted based on their amino acid sequences using for example Jensen-Shannon divergence as we did in our recent approach [27] (see Appendix A.4). The regulatory domain in turn controls the activity of a TF by e.g. ligand binding or phosphorylation and the trans-activation domain is usually characterized by a specific amino acid composition [26].

The human genome consists of around 20000 protein coding genes of which roughly 1500 code for TFs. Considering isoforms that are generated by alternative splicing, the human body contains more than 2900 TFs [25]. However, the number of TFs is much smaller than the number of all genes and consequently the composition of TFs bound to regulatory

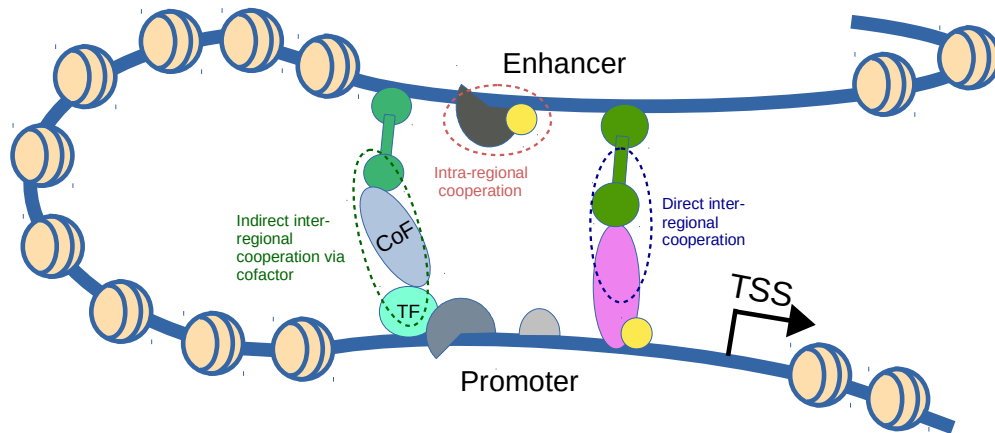




**Figure 2.5.: Mechanisms determining promoter-enhancer interactions.** The pairing of an enhancer to a certain promoter is enabled if a) the bound transcription factors are compatible to each other, b) the spatial constraints allow the contact between the two DNA regions, c) insulator elements do not hinder the pairing and d) the chromatin landscape of the enhancer is accessible. (Based on [21, 24])

elements as well as TF interplay is important in order to provide a proper gene regulation in eukaryotic cells.

Further, TFs in general have an oligomerization domain that allows the direct physical interaction (synergistic or antagonistic) with other TFs. Thereby, TFs form homo- and heterodimers with other TFs, depending on whether the interaction partner is of the same type or not and extending this dimerization process, TFs use to form high order complexes in combination with co-factor proteins. The binding sites of the underlying TFs in turn form clusters on DNA that are known as *cis* regulatory modules. Direct physical cooperations between transcription factors are depicted in Figure 2.6. Regarding a regulatory region, TFs that bind to the *cis* regulatory modules inside that region are interacting with each other. In addition, the TFs that are bound to different regulatory regions can directly physically

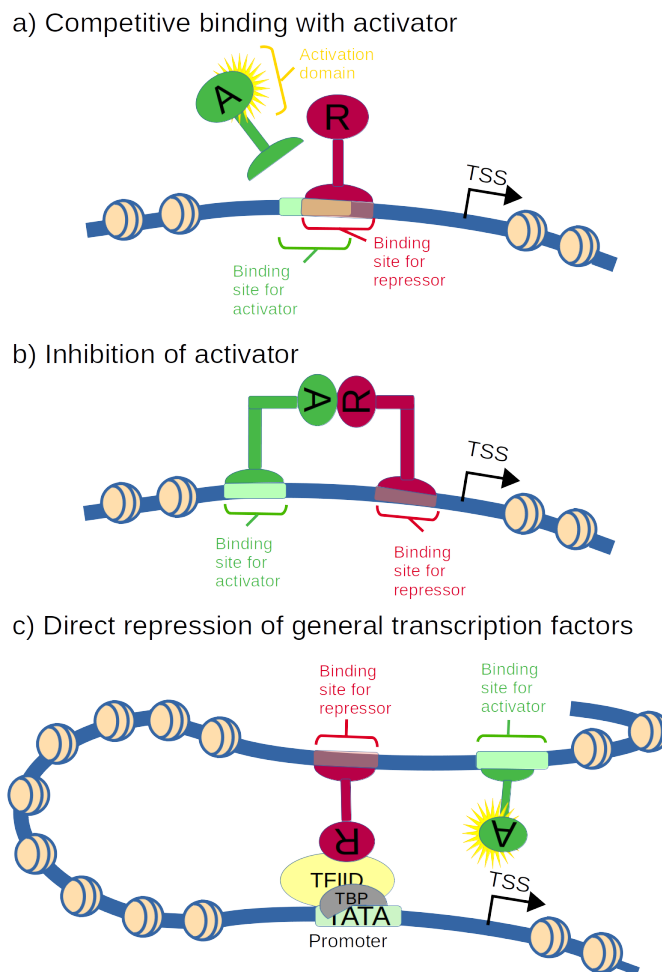


**Figure 2.6.: Physical cooperation strategies of transcription factors.** In order to provide proper gene regulation, transcription factors (TFs) have to cooperate with other TFs or cofactors (CoF) in a synergistic or antagonistic manner. These cooperations can for example take place between TFs that bind next to each other on DNA (intra-sequence cooperations) and TFs that belong to different regulatory sequences (inter-sequence cooperations). The cooperations between TFs of different regulatory regions can be based on direct physical interactions or can be established by cofactors.



**Figure 2.7.: Modular composition of transcription factors.** In general transcription factors consists of all or some of the following domains: DNA binding domain, oligomerization domain, regulatory domain and trans-activation domain.

interact with each other or indirectly via co-factor. These physical cooperations can be synergistic or antagonistic in a way that the effect of activating transcription factors can be strengthened or reduced. For the antagonistic way, transcription factors termed repressors hinder the activity of activating TFs as depicted in Figure 2.8. Regarding one regulatory sequence, repressors can functionally cooperate with the activator by blocking its binding site or physically cooperate by masking its activation domain. In contrast, repressors bound to a distal regulatory region (like enhancer region) can directly or indirectly interact with activating TFs on the promoter [16].



**Figure 2.8.: Strategies of repressing transcription factors.** A transcription factor can fulfill its repressing function by a) blocking the binding site of the transcriptional activator, b) interacting with the activator and thereby covering its activation domain and c) directly repress transcription initiation by interacting with general transcription factors. (Modified from [16, Fig. 17-20])

## 2.2. Experimental methods

The data and methods used in this thesis are based on laboratory experiments. I can not capture all important principles and list a few basic experimental methods for the identification

of transcription factor binding sites as well as for the determination of long range chromatin interactions like promoter-enhancer interactions (PEIs). For more details please have a look at a textbook like [16].

### 2.2.1. Determination of TFBSs

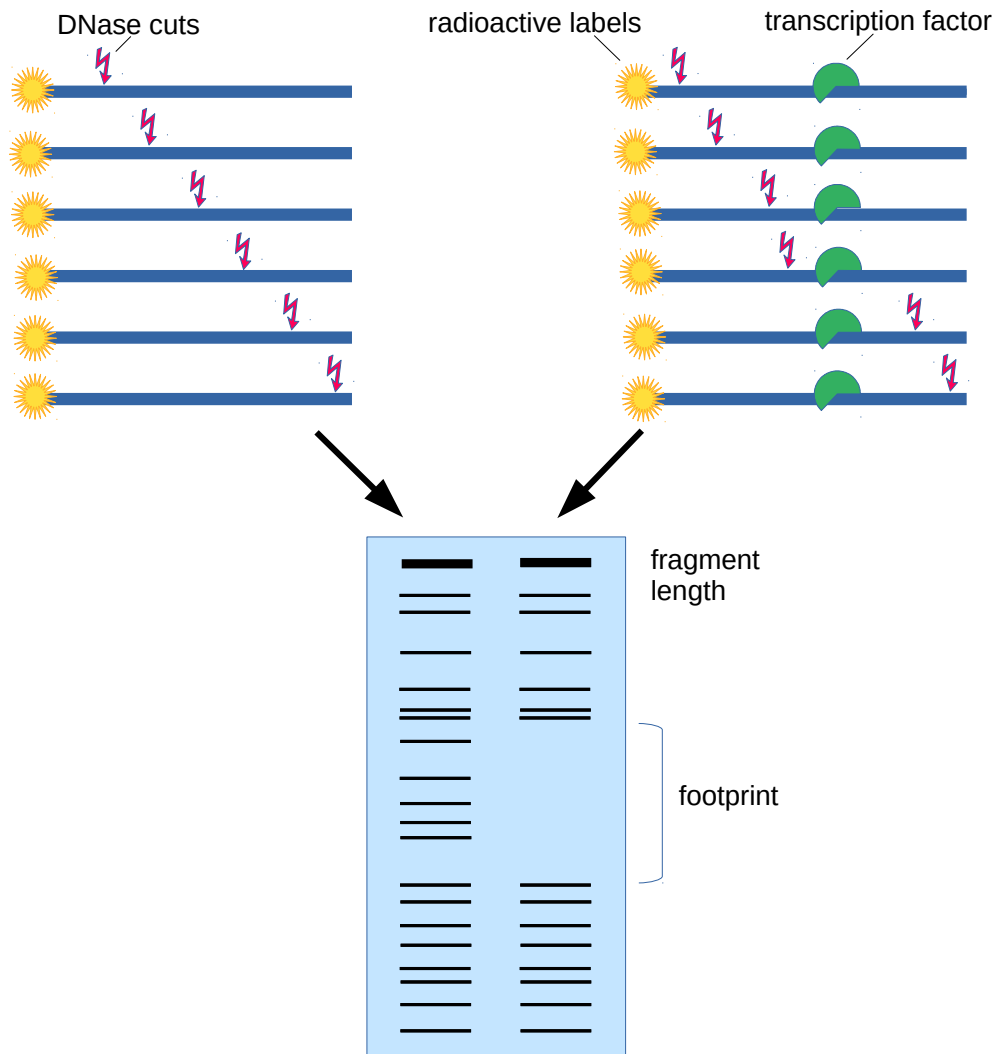
The determination of transcription factor binding sites (TFBSs) is important for the identification of the preferred binding site profiles of a certain factor and in turn for the computational prediction of binding sites in the sequences of interest. In the following, I present exemplarily the nuclease protection footprinting as a method for the determination of protein binding DNA sequences.

**Nuclease protection footprinting** Nucleases are enzymes that cut nucleic acids. A commonly used nuclease in the context of biotechnology is DNase I that cuts one strand of double stranded DNA. If the DNA is bound by proteins the bound regions are protected from a nuclease cleavage. This property is used in the nuclear protection footprinting. One end of the DNA strand is marked (e.g. radioactively labelled) and afterwards the DNA is exposed to a nuclease (e.g. DNase I). The DNA strands are randomly cut by the nuclease and the labeled strands are separated by size in an electrophoresis (see Figure 2.9). The regions bound by the protein cannot be accessed by the nuclease resulting in a lack of DNA strands of particular size (footprint) in the electrophoresis ([16], page 777) .

### 2.2.2. Determination of promoter-enhancer interactions

In the following I give a brief overview of the idea for the determination of long range chromatin interactions like promoter-enhancer interactions (PEIs). According to the present state of the art, such long-range interactions are determined by the chromosome conformation capture.

**Chromosome conformation capture** One of the most popular techniques to determine the topological structure of chromatin is the *chromosome conformation capture* (3C) method. The method identifies long distance DNA regions that are close to each other in the interphase chromatin enhancer-promoter interactions. The general idea of the method is rather simple: In the first step, the chromatin is fixed using e.g. formaldehyde. This chemical introduces covalent bonds (crosslinks) between DNA and the bound proteins. In the next step, the DNA is digested, either by endonucleases like HindIII or BamHI or in a chemical way, followed by the ligation of the free DNA ends. Afterwards, the number of newly created junctions is quantified and statistically evaluated in order to differentiate noise from real signal. Based on the original 3C method, several further methods have been developed which differ in their coverage and general detection aim. In the original 3C method, one can only determine whether two DNA regions of interest are interacting



**Figure 2.9.: Nuclease protection footprinting.** Two sets of the same DNA fragments are which radioactive labels are cut with DNase I. One of the set contains the transcription factor of interest while the other set is not bound by proteins. After DNase cleavage, the DNA fragments are separated according to their length by a gel electrophoresis and the lack of bands (footprint) of the protein containing DNA set indicates the transcription factor binding site. (Figure based on ([16], page 777))

with each other. In 4C, the contacts between the region of interest and genome-wide DNA fragments were determined (*one vs all*), where in 5C genome wide interactions were predicted (*all-vs-all*) [28]. Two newer extensions (*all vs all*) of 3C method are Hi-C and ChIA-Pet and are explained in the following.

**Hi-C** Hi-C is one of the latest extended 3C method. The first steps are (as in the original 3C) the fixation of DNA and DNA cleavage using restriction enzymes. However, before the religation takes place, the ends are filled with biotin-labeled nucleotides and the DNA is purified and sheared and a pull down is performed by using a biotin-antibody. Thereby, only the ligated DNA fragments are considered in the following analysis steps. The pull-down is required, because in contrast to the original 3C method, no primers that could be used for PCR are specified. Afterwards, the reads are mapped back to genomic regions, the number of ligations of long-distance DNA regions are counted and a matrix of fragments is created where an entry refers to the number of counts of the links between the respective fragments. Applying a statistical analysis to this matrix results in the determination of significant genome-wide long distance interacting DNA regions [28].

**ChIA-Pet** A new generation of 3C experiments combines the Hi-C methodology with chromatin immunoprecipitation sequencing (ChIP-Seq). In this method, all potential connections between DNA fragments are predicted in a genome-wide manner (*all-vs-all*) that are bound by a given DNA interacting protein. The overall workflow follows the 3C methodology, fixation of DNA, cleavage and religation. Afterwards, the ligated DNA fragments were pulled down using an antibody against the protein of interest. However, it cannot be determined whether the protein of interest is responsible for the chromatin interaction or just linked to one of the corresponding sequences. The method is restricted in a way that only those DNA fragment connections are determined that are associated with the used protein [28].

## 2.3. Bioinformatic resources

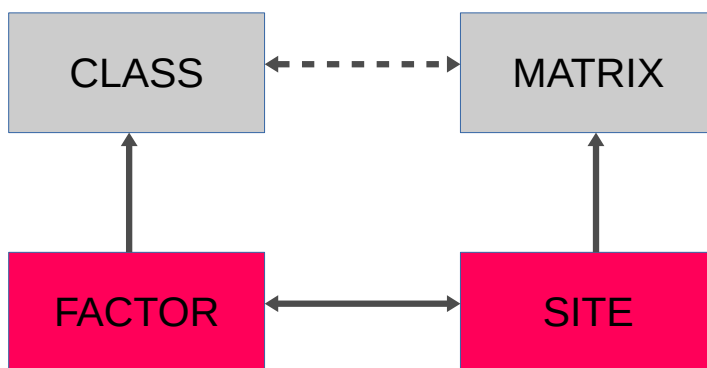
In this section, I will give an overview about the bioinformatic resources used in this thesis. I will start with the databases: TRANSFAC<sup>®</sup>, UCSC Genome Browser, BioGRID, STRING and TRANSCompel. Afterwards, I will shortly present the bioinformatic tools Match<sup>TM</sup> and uShuffle.

### 2.3.1. Bioinformatic data bases

#### 2.3.1.1. TRANSFAC<sup>®</sup>

TRANSFAC<sup>®</sup> has been published by Edgar Wingender for the first time in 1988 [29] and is hosted by the geneXplain company (<http://genexplain.com/>). TRANSFAC<sup>®</sup> is a

database for storing information about eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles. Additionally, for each transcription factor, structural and functional properties are listed and the transcription factors are grouped according to their DNA binding domains in *genus*, *subfamily*, *family*, *class* and *superclass*. The DNA binding sites are experimentally identified and listed with exact genomic position, experimental method and DNA sequence. The DNA sequences are aligned and form the basis for the creation of DNA binding site profiles that are represented as position weight matrices (PWMs) that in turn can be used for the computational prediction of potential transcription factor binding sites (TFBSs) in given regulatory sequences [30].



**Figure 2.10.: The basic structure of TRANSFAC database.** The center of the database is the relation between transcription factor (FACTOR) and its DNA binding site (SITE). On the basis of the binding site sequences, profiles were created (MATRIX) for the prediction of potential binding sites. TFs were grouped according to their binding site domains (CLASS).

The original structure of the database is depicted in Figure 2.10. The center of the database is the relation between a transcription factor and its binding site, stored in the Tables FACTOR and SITE, respectively. The grouping of the factors is placed in Table CLASS and the binding site profiles in Table MATRIX. The number of entries of the main tables in September 2018 is shown in Table 2.1. Up to date, the original TRANSFAC<sup>®</sup> database has been extended by a multitude of additional tables and links to other databases [30].

### Excursus: Position weight matrices

A position weight matrix (PWM) or position specific scoring matrix (PSSM) is a widely accepted model for the representation of biological sequence profiles. It is generally based on sequence alignments and depicts for each motif position the frequency or weight of each letter (i.e. nucleotide or amino acid).

	Pos	A	C	G	T
Sequence 1:	1	3	1	1	1
Sequence 2:	2	0	0	0	5
Sequence 3:	3	5	0	0	0
Sequence 4:	4	1	0	0	4
Sequence 5:	5	4	0	0	1
	6	5	0	0	0
	7	1	4	0	0
	8	0	2	0	3
	9	0	4	0	1

*Creation of a position weight matrix (PWM) on the basis of aligned nucleotide sequences.*

The picture above shows the creation of a position weight matrix (PWM). On the left side, five nucleotide sequences of length nine that belong to a certain sequence profile are aligned to each other. On the right side, the corresponding PWM is shown that stores the frequency of each nucleotide on each alignment position. A generally used way to present PWMs is a logoplot representation.

P0	A	C	G	T	
01	8	0	1	5	W
02	0	3	11	0	G
03	9	2	3	0	A
04	2	0	12	0	G
05	0	0	14	0	G
06	14	0	0	0	A
07	14	0	0	0	A
08	1	3	10	0	G



*TRANSFAC<sup>®</sup> binding site profile V\$PUI\_Q6, as PWM in TRANSFAC format on the left side, including the consensus binding site in the last column and in logoplot representation on the right side.*

### 2.3.1.2. ENCODE

The ENCODE (ENCyclopedia of DNA Elements) project was established by the US National Human Genome Research Institute (NHGRI) in 2003 and was intended to analyze the whole human genome by identifying all functional elements in the underlying DNA regions. Thereby, computational and laboratory scientists work together in the application and analysis of high-throughput experiments for the identification of new structural and functional components encoded in genome sequences. These components include protein-coding genes, non-protein coding genes, transcriptional regulatory elements and regulatory



**Table 2.1.: Latest public statistics of TRANSFAC<sup>®</sup> database in January 2019.** (Source: [http://genexplain.com/wp-content/uploads/2019/01/TRANSFAC\\_statistics\\_2019.1.pdf](http://genexplain.com/wp-content/uploads/2019/01/TRANSFAC_statistics_2019.1.pdf))

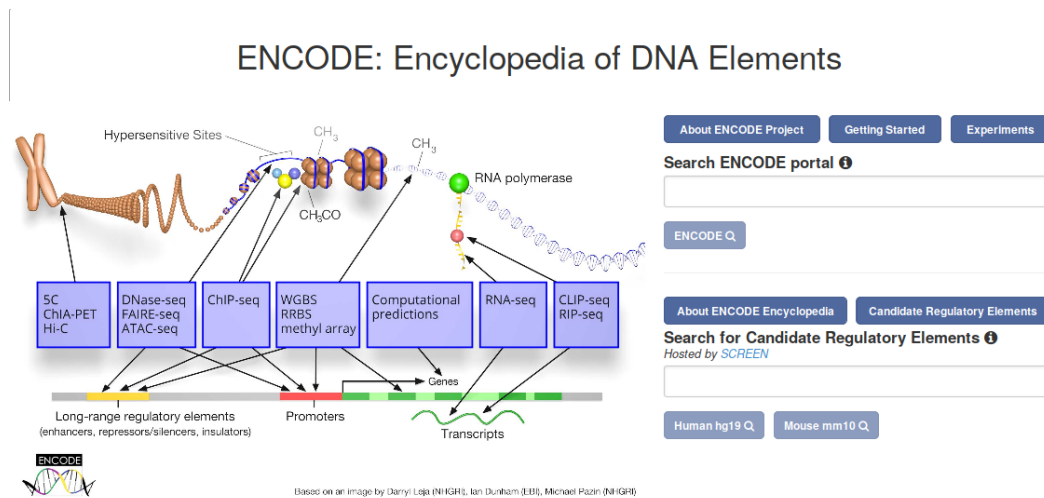
Category	TRANSFAC <sup>®</sup> entries
Factors	47,509
miRNAs1	279
DNA Sites	49,934
mRNA Sites	21,543
Factor-DNA Site Links	67,606
miRNA-mRNA Site Links	57,765
Genes	88,248
ChIP TFBS	83,469,984
Dnase Hypersensitivity Sites	15,376,241
Histone Modification Fragments	1,071,162
DNA Methylation Fragments	51,926
Matrices	8,161
References	37,447

sequence elements monitoring chromosome structure and dynamics [31]. Data provided by ENCODE are freely available (see Figure 2.11) and can also be downloaded in a more structured version by using the UCSC Genome Browser (see Section 2.3.1.3).

### 2.3.1.3. The UCSC Genome Browser

The UCSC Genome Browser (<http://genome.ucsc.edu>) is a public database hosted by the University of California Santa Cruz for genomes and genome annotations of selected species. These annotated data include for example: mRNA, expressed sequence tag (EST) alignments, gene predictions, cross-species homologies and single nucleotide polymorphism [32]. The species range includes vertebrate and non-vertebrate species and some selected model organisms (see Figure 2.12).

The Genome Browser consists of a collection of organism specific databases whereas the tables of each data base are differentiated by *positional tables* and *non-positional tables*. *Positional tables* contain information directly linked to genomic localisations such as gene predictions while *non-positional tables* store information like ID mapping (e.g. which gene ID is linked to which RefSeq ID). These data or a subset of these data can be accessed in text-format using the UCSC Table Browser. A screenshot of the selection interface of the UCSC Table Browser is given in Figure 2.13.



**Figure 2.11.:** Screenshot of the ENCODE search interface. (Source: <https://www.encodeproject.org/>, 17.02.2019)

#### 2.3.1.4. BioGRID

BioGRID (Biological General Repository for Interaction Datasets) has first been published in 2003 (at that time as "The GRID") by Breitkreutz et al. [33] and is hosted by the University of Edinburgh (<https://thebiogrid.org/>). The open source database contains information about protein and genetic interactions, chemical associations and post translational modifications, reported in literature, for the major model organism species, including human. Each interaction is linked to the organism, the experimental method as well as the reference to the original publications. An exemplary search in BioGRID for transcription factor SP1 is shown in Figure 2.14.

**Table 2.2.:** BioGRID statistics of January 2019 for genetical and physical interactions.

Experiment Type	Raw Interactions	Non-Redundant Interactions	Unique Genes	Unique Publications
PHYSICAL	481,059	356,717	22,987	28,528
GENETIC	5,295	5,214	2,192	325
COMBINED	486,354	361,468	23,291	28,654



**Figure 2.12.:** Screenshot of the organism selection menu of the UCSC Genome Browser. The organisms are ordered according to their degree of relationship to human (left side). Scrolling down, more organisms that are less closely related to human are available (right side). (Source: <https://genome.ucsc.edu/cgi-bin/hgGateway>, 14.01.2019)

### 2.3.1.5. STRING

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) published in 2000 by Snel et al. [34] is a database for protein-protein interactions based on experimental validation and computational predictions. These interactions are either direct physical or indirect functional interactions and stem from i) genomic context predictions, ii) high-throughput lab experiments, iii) (conserved) co-expression, vi) automated textmining and v) previous knowledge in databases. In January 2019, the database covers in total 2031 organisms (1678 Bacteria, 238 Eukaryotes, 115 Archaea) and 9,643,763 proteins that share 1,380,838,440 interactions which are in turn grouped by confidence level. The database is freely available at <https://string-db.org/>. Exemplarily, the STRING output for protein SP1 is shown in Figure 2.15.

**Table Browser**

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence in this form, and the [User's Guide](#) for general information and sample queries. For more complex queries, you may want to use [Galaxy](#) or our [public M](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions as page.

**clade:** Mammal  **genome:** Human  **assembly:** Dec. 2013 (GRCh38/hg38)

**group:** Genes and Gene Predictions  **track:** GENCODE v29

**table:** knownGene

**region:**  genome  position

**identifiers (names/accessions):**

**filter:**

**intersection:**

**correlation:**

**output format:** all fields from selected table  Send output to  Galaxy  GREAT  GenomeSpace

**output file:**  (leave blank to keep output in browser)

**file type returned:**  plain text  gzip compressed


**Figure 2.13.: Screenshot of the interface of the UCSC Table Browser.** The user can paste identifiers (e.g. gene names) in order to only access the records of interest. (Source: [https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=706768895\\_h7nSGXKTqUH0tkwaSq3EULRRZ4Kt](https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=706768895_h7nSGXKTqUH0tkwaSq3EULRRZ4Kt), 14.01.2019)

### 2.3.1.6. TRANSCompel

TRANSCompel<sup>®</sup> is a complement of the TRANSFAC<sup>®</sup> database published in 2002 by Kel-Margoulis et al. [14] and is hosted by the geneXplain company (<http://genexplain.com/>). TRANSCompel<sup>®</sup> contains experimentally verified data about eukaryotic composite regulatory elements (CE), closely linked transcription factor binding sites representing small combinatorial regulatory units, with experimental evidence. The CEs are classified according to their two constituents (factor1/factor2) in i) inducible/inducible, ii) inducible/constitutive, iii) tissue-restricted/ubiquitous, iv) inducible/tissue-restricted and v) tissue-restricted/tissue-restricted. The latest statistics for the actual TRANSCompel<sup>®</sup> version (January 2019) is shown in Table 2.3.

**Table 2.3.: Statistics of TRANSCompel<sup>®</sup> database of January 2019.**

Composite elements	Genes	Evidence codes	References
593	402	2,181	661

**BioGRID 3.5** home help wiki tools contribute stats downloads partners about us | 

**Result Summary** Gene / Identifier Search: SP1  All Organisms

**SP1** *Homo sapiens*

Sp1 transcription factor

UBI SUMO


GO Process (10) GO Function (16) GO Component (2)

EXTERNAL DATABASE LINKOUTS  
[BioGRID ORCS](#) | [VEGA](#) | [HGNC](#) | [OMIM](#) | [Entrez Gene](#) | [RefSeq](#) | [UniprotKB](#) | [Ensembl](#) | [HPRD](#)

[Download 578 Published Interactions For This Protein](#)

**Stats & Options**

**Current Statistics** Publications: 243  
 High Throughput: 71 (12%) 575 Physical Interactions 504 (88%)  
 Low Throughput: 0 (0%) 3 Genetic Interactions 3 (100%)

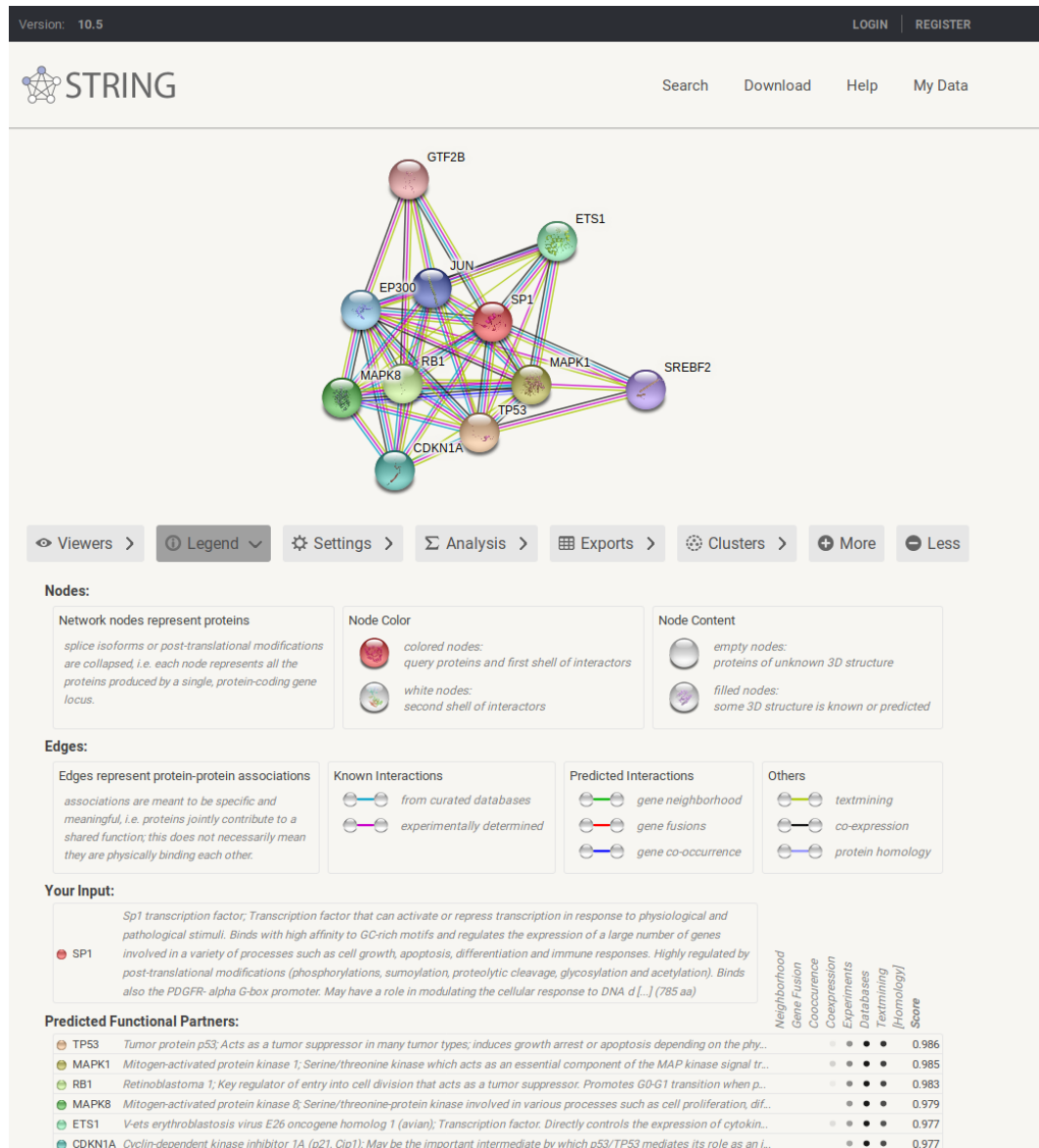
**Search Filters** Customize how your results are displayed...  
 No Filter: Show All Associations 

Switch View: **Interactors (235)** Interactions (578) Network PTM Sites (6)

Displaying 235 total unique interactors  
 Sort By: **[Evidence]** [Alphabetical]

<b>HDAC1</b>   RP4-811H24.2, GON-10, HD1, RPD3, RPD3L1 histone deacetylase 1 UBI SUMO	28 <a href="#">[details]</a>
<b>EP300</b>   RP1-85F18.1, KAT3B, RSTS2, p300 E1A binding protein p300 UBI SUMO	22 <a href="#">[details]</a>
<b>TP53</b>   BCC7, LFS1, P53, TRP53 tumor protein p53 UBI NEDD FAT10 SUMO	14 <a href="#">[details]</a>
<b>ESR1</b>   RP1-130E4.1, ER, ESR, ESRA, ESTRR, Era, NR3A1 estrogen receptor 1 UBI SUMO	13 <a href="#">[details]</a>
<b>HDAC2</b>   HD2, RPD3, YAF1 histone deacetylase 2 UBI NEDD SUMO	10 <a href="#">[details]</a>

**Figure 2.14.:** Screenshot of the result page for SP1 protein interactions in human. Listed are interaction partners of SP1 with the experimental methods and the number of evidences/publications for the respective interaction. (Source: <https://thebiogrid.org/>, 14.01.2019)



**Figure 2.15.:** Screenshot of STRING result page of SP1 protein in human. Depicted is the interaction network of SP1 and its interaction partners in combination with the network legend and the detailed list of interaction partners below. (Source: <https://string-db.org/>, 14.01.2019)

## 2.3.2. Bioinformatic tools

### 2.3.2.1. Match<sup>TM</sup>

Match<sup>TM</sup> [35] is a tool for the prediction of potential transcription factor binding sites in regulatory DNA sequences on the basis of position weight matrices (PWMs). The algorithm scans for each PWM the input sequences and determines the quality of potential PWM-sequence matches with two scores: i) matrix similarity score (MSS) and ii) core similarity score (CSS). While the MSS considers the entire PWM length  $L$ , the CSS only uses the core of a PWM where the core of a PWM is defined as the five most conserved positions. Both scores are in the range between 0.0 and 1.0 (where 1.0 indicates a perfect PWM-sequence match) and are calculated as follows

$$SS = \frac{Current - Min}{Max - Min}, \quad (2.3.1)$$

where SS is short for MSS or CSS. The value *Current* is calculated as follows

$$Current = \sum_{i=1}^L I(i) f_{i,b_i}, \quad (2.3.2)$$

where  $f_{i,b_i}$  is the frequency of nucleotide  $b_i$  at position  $i$  of the PWM for  $b_i \in \{A, C, G, T\}$ . Further *Min* is defined as

$$Min = \sum_{i=1}^L I(i) f_i^{min}, \quad (2.3.3)$$

where  $f_i^{min}$  is the frequency of the rarest nucleotide at position  $i$  of the PWM.

In the same way, *Max* is defined as:

$$Max = \sum_{i=1}^L I(i) f_i^{max}, \quad (2.3.4)$$

where  $f_i^{max}$  is the frequency of the dominating nucleotide at position  $i$  of the PWM.

For the calculation of *Current*, *Min* and *Max* an information vector is defined in the following way:

$$I(i) = \sum_{b_i \in A, G, C, T} f_{i,b_i} \ln(4f_{i,b_i}), \quad (2.3.5)$$

for  $i = 1, 2, \dots, L$ .

In order to evaluate the significance of a match, the algorithm uses pre-specified *cut-off* values for each PWM of TRANSFAC database like: i) minimizing the number of false positive matches (*minFP*), ii) minimizing the number of false negative matches (*minFN*) and iii) minimizing the sum of false negative and false positive matches (*minSUM*). All matches

that exceed the specified threshold are listed in the Match<sup>TM</sup> result file as exemplarily shown in Figure 2.16.

```

1 Search for sites by WeightMatrix library: data/matrix.dat
2 Sequence file: sequences.fasta
3 Site selection profile: prfs/vertebrate_non_redundant_minFP.prf Matrices of vertebrate non-redundant (VNR) with cut offs
4 to minimize false positive rates.
5
6
7 Inspecting sequence ID   sequence_1
8
9 V$CREBP1_01           |      140 (+) | 0.766 | 0.849 | ATACGtaa
10 V$CREBP1_01          |      1478 (+) | 0.743 | 0.714 | TGACAtac
11 V$CREBP1_01          |      1753 (-) | 0.766 | 0.723 | ataCTTAA
12 V$CREBP1_01          |      2015 (+) | 0.613 | 0.750 | TTCCAtaa
13 V$CREBP1_01          |      2892 (-) | 0.766 | 0.849 | ttaCCTAA
14 V$CREBP1_01          |      3081 (+) | 0.597 | 0.740 | TTATTtaa
15 V$CREBP1_01          |      3081 (-) | 0.613 | 0.750 | ttaTTTAA
16 V$CREBP1_01          |      4372 (-) | 0.717 | 0.817 | ttaTGTTA
17 V$CREBP1_01          |      4432 (-) | 0.831 | 0.740 | ttTAGTAA
18 V$CREBP1_01          |      4733 (-) | 0.597 | 0.740 | ttaGGAAA
19 V$DELTAEF1_01        |      1166 (-) | 1.000 | 0.980 | ccaAGGTgggc
20 V$DELTAEF1_01        |      1860 (+) | 1.000 | 0.978 | atgCACCTaga
21 V$DELTAEF1_01        |      3769 (+) | 1.000 | 0.983 | atTCACCTgtg
22 V$CDPCR1_01          |           2 (-) | 0.865 | 0.801 | aataTTGATa

```

**Figure 2.16.: Example of a MATCH<sup>TM</sup> output.** The first column gives the identifier of the TRANSFAC PWM, followed by the first sequence position and strand where the match has been detected. Column three gives the core similarity score (CSS) while the matrix similarity score (MSS) is in column four. The last column contains the matching sequence.

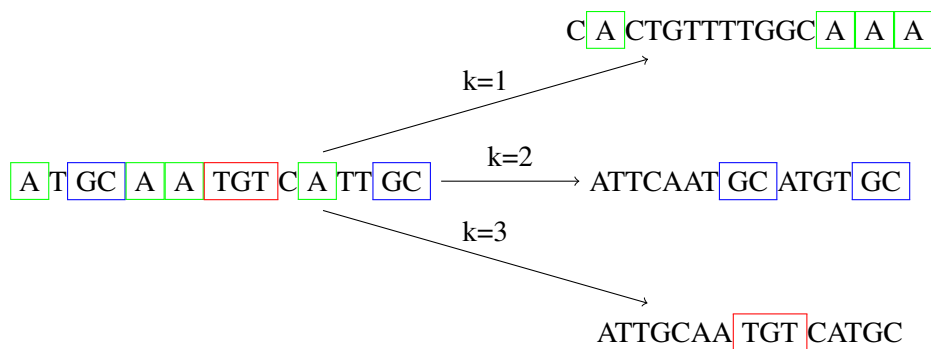
### 2.3.2.2. uShuffle

The uShuffle [36] algorithm has been developed in 2008 by Jiang et al. and is a powerful tool for randomly shuffling biological sequences by preserving the exact  $k$ -mers counts where  $k$ -mers are letter combinations of length  $k$ . Implementing the shuffling process, the uShuffle algorithm uses the Euler algorithm in combination with Wilson's algorithm for the generation of the arborescence. The Euler algorithm is designed for the random generation of uniform shuffled sequences preserving the  $k$ -mers counts and is based on a graph theoretical approach. For details see [36].



### Example: Shuffling of sequences by preserving the $k$ -mers counts

The uShuffle algorithm shuffles sequences by preserving the  $k$ -let counts. Considering a sequence of nucleotides of length 15. Setting  $k = 1$ , the algorithm permutes all nucleotides in the sequence where the frequency of the single nucleotides is maintained. Setting  $k = 2$ , the number of dinucleotides as well as the number of single nucleotides is preserved. Setting  $k = 3$ , the frequency of trinucleotides and again the number of single nucleotides is maintained in input and shuffled sequence.



*Input (left) and output (right) sequences of the uShuffle algorithm for different  $k$ -mers.*

Exemplarily, some  $k$ -mers are marked in color. Considering  $k = 1$  the number of single nucleotides is remaining, e.g. four “A” in both sequences. Setting  $k = 2$ , the number of dinucleotides is preserved as indicated by two “GC”s in the input and shuffled sequence and for  $k = 3$  the number of single and trinucleotides is maintained e.g. one “TGT” in both sequences.



## 3. Theoretical background

In this chapter I will focus on the concept of information theory. I will start with the Shannon Entropy and general information theoretic measures. Further, I will show mutual information measures for two and three random variables like pointwise mutual information, multivariate mutual information, conditional mutual information as well as dual total correlation.

### 3.1. Information theory

Information theory is the mathematical approach to quantify the amount of information of an outcome, a system or a process. In 1940s, the first attempt of this quantification was done in the context of channel capacity by Shannon, who discovered that random processes have an irreducible complexity which he termed *entropy* [37]. Later, a multitude of different measures and quantities arises for the determination of a certain information content considering different numbers of variables in different contexts (such as pairwise and multivariate mutual information, conditional mutual information or dual total correlation).

This section gives a general overview about entropy, mutual information and related measures. The content as well as the notation of this chapter is based on [37].

#### 3.1.1. Entropy

The entropy of a discrete random variable  $X$  of alphabet  $\mathfrak{X} = \{x_1, x_2, \dots, x_n\}$  (where  $|\mathfrak{X}| = n$ ) is a quantitative measure of its uncertainty depending on the probability mass function  $p(x = X) = \{p(x_1), p(x_2), \dots, p(x_n)\}$  where  $\sum_{i=1}^n p(x_i) = 1$ .

**Definition 3.1 (Entropy)** *Let  $X$  be a discrete random variable of alphabet  $\mathfrak{X}$  with probability mass function  $p(x)$ . The entropy  $\mathbb{H}$  of  $X$  is defined by*

$$\mathbb{H}(X) = - \sum_{x \in \mathfrak{X}} p(x) \log p(x). \quad (3.1.1)$$

Since  $\lim_{x \rightarrow 0} x \log x = 0$ , the general convention is to set  $0 \log 0 = 0$ . The base of the logarithm  $b$  can be used to scale the entropy. In case of  $b = 2$ , the unit of the entropy is one bit. In

this thesis, I take the logarithm to base 2 and, thus,  $\log$  stands for  $\log_2$ .  $\mathbb{H}(X) = 0$  if there is no uncertainty in the occurrence of  $X$ , indicating that  $X$  always shows the same value or the same letter  $x$  and  $p(x) = 1$ . The maximal entropy is reached if  $X$  is uniform distributed over all letters of  $\mathfrak{X}$ .

**Example: Calculation of entropy**

Consider the following sequence of outcomes of random variable  $X$ :

*a t t a c g a a*

The aim is to determine the entropy  $\mathbb{H}(X)$  for  $X$  of alphabet  $\mathfrak{X} = \{a, c, g, t\}$ . First of all, the marginal probabilities for each letter in  $\mathfrak{X}$  have to be determined:

$$p(a) = \frac{4}{8} \quad p(c) = \frac{1}{8}$$

$$p(g) = \frac{1}{8} \quad p(t) = \frac{2}{8}$$

The entropy is then calculated as follows:

$$\begin{aligned} \mathbb{H}(X) &= -\sum_{i=1}^4 p(x_i) \log_2(p(x_i)) \\ &= -(p(a) \log_2(p(a)) + p(c) \log_2(p(c)) + p(g) \log_2(p(g)) + p(t) \log_2(p(t))) \\ &= -\left(\frac{4}{8} \cdot \log_2\left(\frac{4}{8}\right) + \frac{1}{8} \cdot \log_2\left(\frac{1}{8}\right) + \frac{1}{8} \cdot \log_2\left(\frac{1}{8}\right) + \frac{2}{8} \cdot \log_2\left(\frac{2}{8}\right)\right) \\ &= -(0.5 + 0.375 + 0.375 + 0.5) \\ &= -1.75 \end{aligned}$$

Finally, the entropy  $\mathbb{H}(X) = 1.75$  bits.

The formula of entropy can be extended for two random variables resulting in the joint entropy. The joint entropy of two random variables  $X$  and  $Y$  of alphabet  $\mathfrak{X}$  and  $\mathfrak{Y}$ , respectively, is defined in a similar way as the single entropy by using the joint probability distribution  $p(x, y)$  and using  $\mathfrak{X} \times \mathfrak{Y}$  as a kind of extended alphabet.

**Definition 3.2 (Joint entropy)** Let  $X$  and  $Y$  be two discrete random variables with joint probability mass function  $p(x,y)$ . The joint entropy is then defined as

$$\mathbb{H}(X,Y) = - \sum_{x \in \mathfrak{X}} \sum_{y \in \mathfrak{Y}} p(x,y) \log p(x,y). \quad (3.1.2)$$

The conditional entropy  $\mathbb{H}(Y|X)$  describes the uncertainty of a discrete random variable  $Y$  given the knowledge of random variable  $X$ .

**Definition 3.3 (Conditional Entropy)** The conditional entropy of two discrete random variables  $X$  and  $Y$  is given as

$$\mathbb{H}(Y|X) = - \sum_{x \in \mathfrak{X}} \sum_{y \in \mathfrak{Y}} p(x,y) \log p(y|x). \quad (3.1.3)$$

The relation of the conditional entropy, joint entropy as well as the marginal entropies of two random variables  $X$  and  $Y$  is expressed by the theorem of *chain rule* (for proof see [37]) as

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X) \quad (3.1.4)$$

Having a closer look to Figure 3.1 the following properties of entropy can be deduced:

- $\mathbb{H}(X) \geq 0$
- $\mathbb{H}(X,Y) = \mathbb{H}(Y,X)$
- $\mathbb{H}(X|Y) \neq \mathbb{H}(Y|X)$  with equality if and only if,  $\mathbb{H}(X) = \mathbb{H}(Y)$
- $\mathbb{H}(X,Y) \leq \mathbb{H}(X) + \mathbb{H}(Y)$
- $\mathbb{H}(X,Y) \geq \max\{\mathbb{H}(X), \mathbb{H}(Y)\}$  with equality if one is enclosed in the other.

### 3.1.2. Mutual Information

The mutual information of two random variables  $X$  and  $Y$  is a measure of the information that one random variable contains about the other. It can also be described as the reduction of uncertainty of a random variable due to the knowledge of the other.

**Definition 3.4 (Mutual information)** Let  $X$  and  $Y$  be two discrete random variables with marginal probability mass functions  $p(x)$  and  $p(y)$ , respectively. Considering the joint distribution  $p(x,y)$  the mutual information  $\mathbb{I}(X,Y)$  of  $X$  and  $Y$  is defined as

$$\mathbb{I}(X;Y) = \sum_{x \in \mathfrak{X}} \sum_{y \in \mathfrak{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (3.1.5)$$

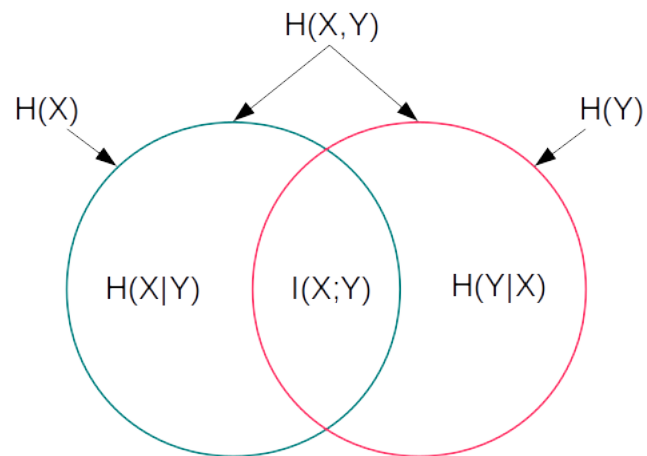
The mutual information is strongly related to the entropies of the random variables (see Figure 3.1) and due to symmetry, it can be expressed as

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) \quad (3.1.6)$$

and

$$\mathbb{I}(X;Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \quad (3.1.7)$$

It can easily be seen that  $0 \leq \mathbb{I}(X;Y) \leq \min\{\mathbb{H}(X), \mathbb{H}(Y)\}$ .



**Figure 3.1.: Relation between entropy and mutual information.**

**Example: Calculation of mutual information**

Consider the following example:

$$\begin{array}{cccccc} a & a & c & c & a & a \\ t & t & g & g & t & t \end{array}$$

The aim is to determine the mutual information  $\mathbb{I}(X, Y)$  for random variables  $X$  and  $Y$  of alphabets  $\mathcal{X} = \{a, c\}$  and  $\mathcal{Y} = \{g, t\}$ , respectively. First of all, the marginal probabilities as well as the pairwise probabilities have to be determined :

Marginal probabilities of  $X$ :

$$p(a) = \frac{4}{6}$$

$$p(c) = \frac{2}{6}$$

Joint probabilities:

$$p(a, g) = 0 \quad p(c, g) = \frac{2}{6}$$

$$p(a, t) = \frac{4}{6} \quad p(c, t) = 0$$

Marginal probabilities of  $Y$ :

$$p(g) = \frac{2}{6}$$

$$p(t) = \frac{4}{6}$$

The mutual information  $\mathbb{I}(X, Y)$  between  $X$  and  $Y$  is then calculated as follows:

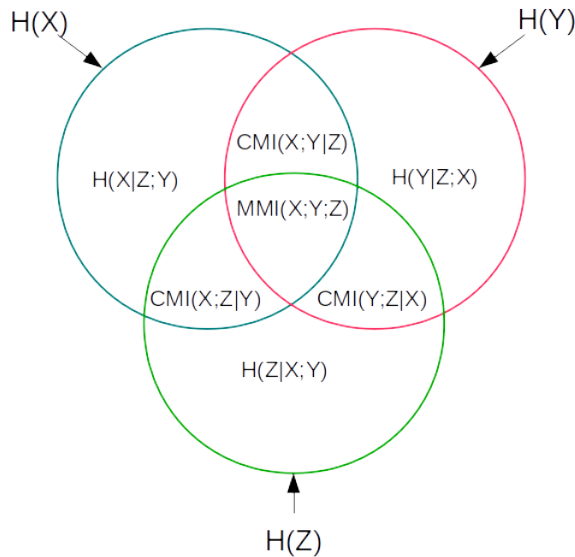
$$\begin{aligned} \mathbb{I}(X; Y) &= \sum_{i=1}^2 \sum_{j=1}^2 p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \\ &= p(a, g) \log_2 \frac{p(a, g)}{p(a)p(g)} + p(a, t) \log_2 \frac{p(a, t)}{p(a)p(t)} + p(c, g) \log_2 \frac{p(c, g)}{p(c)p(g)} + p(c, t) \log_2 \frac{p(c, t)}{p(c)p(t)} \\ &= 0 + \frac{4}{6} \cdot \log_2 \left( \frac{4/6}{4/6 \cdot 4/6} \right) + \frac{2}{6} \cdot \log_2 \left( \frac{2/6}{2/6 \cdot 2/6} \right) + 0 \\ &= 0.389975 + 0.5283208 \\ &= 0.9182958 \end{aligned}$$

Finally, the mutual information  $\mathbb{I}(X, Y) = 0.918$ .

### 3.1.3. Multivariate mutual information

The classical information theoretical approaches can be extended for systems with more than two random variables. Thereby, several different multivariate mutual information theoretic measures have been established that try to analyze the dependency and relationship of a multitude of random variables. In this work, I will only focus on three random variables. It has to be noted, that naming of the different measures is not consistent throughout the literature. In order to avoid confusion, I listed for each measure the mathematical notation and formula.

Given three discrete random variables  $X$ ,  $Y$  and  $Z$  of alphabets  $\mathfrak{X} = \{x_1, x_2, \dots, x_n\}$ ,  $\mathfrak{Y} = \{y_1, y_2, \dots, y_m\}$  and  $\mathfrak{Z} = \{z_1, z_2, \dots, z_l\}$  with length  $n$ ,  $m$  and  $l$ , respectively.



**Figure 3.2.: Relation between entropy, mutual information and multivariate mutual information** for three random variables  $X, Y$  and  $Z$ .

The easiest way to calculate the interaction strength of three random variables  $X$ ,  $Y$  and  $Z$  is to use the pairwise mutual information  $\mathbb{I}(X, S)$  of one variable  $X$  and the grouping  $S = Y, Z$  of  $Y$  and  $Z$  (see Figure 3.3 a)). This can also be expressed as  $\mathbb{JMI}(X; Y, Z)$  and is termed *joint mutual information* in the following chapters. However, by grouping the variables in  $S$  and considering  $S$  as one random variable, the impact of the individual variables cannot be differentiated from those of the others [38]. The properties are the same as for the pairwise mutual information by considering two single random variables.



**Definition 3.5 (Joint mutual information)** *The mutual information  $\mathbb{J}\text{MII}$  of a pair of random variables  $X$  and  $Y$  with a third random variable  $Z$  is defined as*

$$\mathbb{J}\text{MII}(X, Y; Z) = \mathbb{H}(X, Y) + \mathbb{H}(Z) - \mathbb{H}(X, Y, Z) \quad (3.1.8)$$

and can also be expressed by probability mass functions as

$$\mathbb{J}\text{MII}(X, Y; Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \log \frac{p(x, y, z)}{p(x, y)p(z)} \quad (3.1.9)$$

The conditional mutual information  $\mathbb{C}\text{MII}(X; Y|Z)$  describes the reduction of uncertainty of  $X$  having the knowledge of  $Y$  when  $Z$  is given. It is depicted in Figure 3.3 b).

**Definition 3.6 (Conditional mutual information)** *The conditional mutual information  $\mathbb{C}\text{MII}$  of discrete random variables  $X$  and  $Y$  given  $Z$  is defined by the probability mass functions as*

$$\mathbb{C}\text{MII}(X; Y|Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \quad (3.1.10)$$

and can also be expressed by the entropies as.

$$\mathbb{C}\text{MII}(X; Y|Z) = \mathbb{H}(X|Z) - \mathbb{H}(X|Y, Z) \quad (3.1.11)$$

An important property of conditional mutual information is:

- $\mathbb{C}\text{MII}(X; Y|Z) \geq 0$  with equality if and only if  $X$  and  $Y$  are conditional independent given  $Z$ .

The multivariate mutual information  $\mathbb{M}\text{MII}$  of three random variables  $X$ ,  $Y$  and  $Z$  is defined as the intersection of all pairwise mutual information as shown in Figure 3.3 c).

**Definition 3.7 (Multivariate mutual information)** *The multivariate mutual information  $\mathbb{M}\text{MII}$  of three random variables  $X$ ,  $Y$  and  $Z$  can be expressed by pairwise mutual information and conditional mutual information as*

$$\begin{aligned} \mathbb{M}\text{MII}(X; Y; Z) &= \mathbb{I}(X; Y) - \mathbb{C}\text{MII}(X; Y|Z) \\ &= \mathbb{I}(X; Z) - \mathbb{C}\text{MII}(X; Z|Y) \\ &= \mathbb{I}(Y; Z) - \mathbb{C}\text{MII}(Y; Z|X) \end{aligned} \quad (3.1.12)$$

Properties of multivariate mutual information

- Symmetry with regard to  $X$ ,  $Y$  and  $Z$
- Bounds:

$$-\min\{\text{CMI}(X;Y|Z), \text{CMI}(X;Z|Y), \text{CMI}(Z;Y|X)\} \leq \text{MMI}(X;Y;Z) \leq \min\{\mathbb{I}(X;Y), \mathbb{I}(Y;Z), \mathbb{I}(X;Z)\}$$

- If  $X$ ,  $Y$  and  $Z$  form a Markov chain  $X \rightarrow Y \rightarrow Z$  then  $\text{MMI}(X;Y;Z) \geq 0$ . This follows from the property of a Markov chain that  $\text{CMI}(X;Y|Z) \leq \mathbb{I}(X;Y)$  (for proof see [37])

In contrast to the pairwise mutual information, the multivariate mutual information can become negative, for example, if  $X$  and  $Y$  are independent of each other ( $\mathbb{I}(X;Y) = 0$ ), but become dependent by the knowledge of  $Z$  ( $\mathbb{I}(X;Y|Z) \geq 0$ ).

The dual total correlation has first been described in 1978 by Han [39] and considering three random variables  $X, Y$  and  $Z$  it is the union of their pairwise mutual information (see Figure 3.3 d)).

**Definition 3.8 (Dual total correlation)** *The dual total correlation  $\mathbb{DTC}$  of three random variables  $X$ ,  $Y$  and  $Z$  is defined as*

$$\mathbb{DTC}(X, Y, Z) = \mathbb{I}(X;Z|Y) + \mathbb{I}(Y;Z|X) + \mathbb{I}(X;Y) \quad (3.1.13)$$

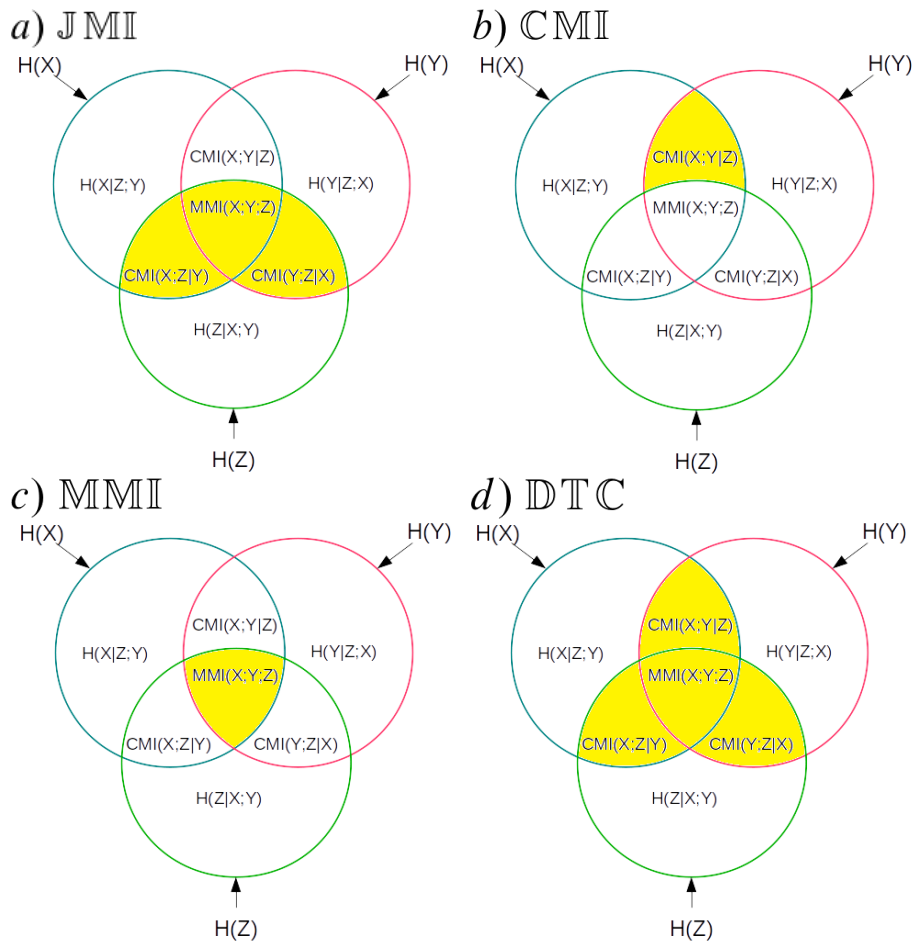
Properties of dual total correlation:

- $\mathbb{DTC}(X, Y, Z) \geq 0$  followed by definition
- $\mathbb{DTC}(X, Y, Z) \leq \mathbb{H}(X, Y, Z)$
- Symmetric with regard to  $X$ ,  $Y$  and  $Z$

### 3.1.4. Pointwise mutual information

The pointwise mutual information (PMI) is a measure for the association strength of two outcomes  $X = x$  and  $Y = y$  where  $X$  and  $Y$  are two discrete random variables. In 1990 Church and Hanks [40] used it for the first time in the field of linguistics as psycholinguistic association score. Since then, the PMI has become more and more popular in linguistics for the determination of word collocations [41] as well as document summarizing processes [42]. Word collocations are sequences of words (i.e. pairs of words) that co-occur more often than expected by chance in a document.

An example for a word collocation is “major problem”. The words “major” and “problem” can often be found together in a text or a conversation and are therefore in a statistical dependence of each other. In contrast, the word combination “blue problem” is not common and rather untypical. The probability to find this combination is at most equal or less than expected by pure chance and thus, these two words do not form a collocation.



**Figure 3.3:** Information theory measures for three random variables. a) joint mutual information (JMI), b) conditional mutual information (CMI), c) multivariate mutual information (MMI) and d) dual total correlation (DTC).

**Definition 3.9 (Pointwise mutual information)** *The pointwise mutual information  $\text{PMI}$  of two discrete random variables  $x \in X$  and  $y \in Y$  is given by*

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad (3.1.14)$$

where  $p(x, y)$  is the joint distribution of  $x$  and  $y$  and  $p(x)$  and  $p(y)$  are the marginal probabilities, respectively.

Properties of  $\mathbb{P}\text{MII}$ :

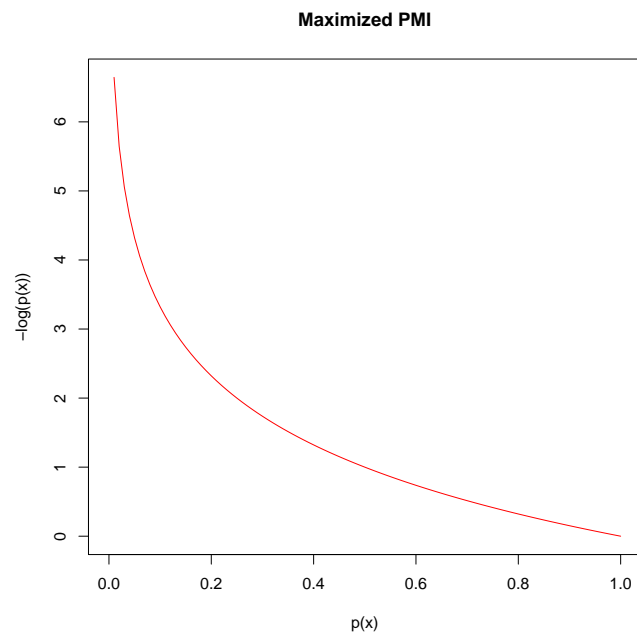
- $\mathbb{P}\text{MII}(x, y) = 0 \leftrightarrow p(x, y) = p(x)p(y)$
- $\mathbb{P}\text{MII}(x, y) > 0$  if  $x$  and  $y$  are dependant in case  $X = x$  and  $Y = y$
- $\mathbb{P}\text{MII}(x, y) < 0$  if  $X$  and  $Y$  are independent in case  $X = x$  and  $Y = y$
- $-\infty \leq \mathbb{P}\text{MII}(x, y) \leq \min[-\log(p(x)), -\log(p(y))]$

Having a closer look to Formula 3.1.14 it becomes clear that  $\mathbb{P}\text{MII}(x, y)$  is equal to zero if  $x$  and  $y$  are independent of each other.  $\mathbb{P}\text{MII}(x, y)$  is in a positive range, if they co-occur more often than expected by pure chance and it is maximized if  $x$  and  $y$  are perfectly associated. The differences between  $\mathbb{P}\text{MII}$  and  $\mathbb{I}$  are depicted in Table 3.1

**Table 3.1.: Difference between pointwise mutual information  $\mathbb{P}\text{MII}$  and mutual information  $\mathbb{I}$**

Mutual information	Pointwise mutual information
$\mathbb{I}(X, Y) \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$	$\mathbb{P}\text{MII}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$
Refers to the average of all events	Refers to a single event
Measure for the amount of information one random variable tells about another random variable	Measure of association of how much one outcome tells about another outcome

A specificity of  $\mathbb{P}\text{MII}$  is its susceptibility to low number counts which is also illustrated in the Example below but can also be seen easily in the formula: if two outcomes  $x$  and  $y$  occur only once in the system and this time, they occur together than  $p(x) = p(y) = p(x, y)$  following that  $\mathbb{P}\text{MII}(x, y) = -\log(p(x))$  indicating a perfect association [41]. A closer look to Figure 3.4 indicates that the function of  $-\log(p(x))$  is decreasing by increasing  $p(x)$ . Thus, by considering two perfectly associated outcomes, the one with less occurrences has a higher  $\mathbb{P}\text{MII}$  value.



**Figure 3.4.: Maximized PMI in dependence of occurrence probability  $p(x)$  of  $x$ .** A low occurrence probability results in a high PMI-value while outcomes with a high occurrence probability in the system gets comparative small PMI-values.

**Example: Calculation of pointwise mutual information**

Considering the following example:

*aa ab bc ab ba bb ba*

The aim is to determine the pointwise mutual information  $\text{PMI}$  for each pair of letters. First of all, the marginal probabilities of each letter as well as the pairwise probabilities of all letter pairs have to be determined :

Marginal probabilities:

$$p(a) = \frac{6}{14}$$

$$p(b) = \frac{7}{14}$$

$$p(c) = \frac{1}{14}$$

Joint probabilities:

$$p(a,a) = \frac{1}{7} \quad p(b,c) = \frac{1}{7}$$

$$p(a,b) = \frac{4}{7} \quad p(c,c) = 0$$

$$p(b,b) = \frac{1}{7}$$

The pointwise mutual information of  $ab$  is then calculated as follows:

$$\text{PMI}(a,b) = \log_2 \frac{p(a,b)}{p(a)p(b)} = \log_2 \frac{4/7}{6/14 \cdot 7/14} = \log_2 2.67 = 1.415$$

In the same way the  $\text{PMI}$  for all other letter pairs is calculated leading to the following results:

$$\text{PMI}(a,a) = -0.36 \quad \text{PMI}(b,b) = -0.807$$

$$\text{PMI}(b,c) = 2 \quad \text{PMI}(c,c) = 0$$

In consideration of entropy calculation, I define  $\log_2 = 0$ . This example also shows the overestimation of low number counts by  $\text{PMI}$  since  $c$  occurs only one time and  $\text{PMI}(b,c) > \text{PMI}(a,b)$ , although  $a$  and  $b$  are much more co-occurring.

## 4. Information theoretical approaches for the analysis of cooperating TFs

In this chapter I will present two information theoretic approaches developed in this thesis for the identification of cooperating TFs based on their binding site distributions. In the first part, I present a method based on pointwise mutual information (PMI) for the identification of co-occurring TFBSs in a regulatory sequence (intra- sequence TFBS collaborations). In the second part of the chapter, I use different multivariate information theoretical measures for the identification of associated TFBSs in promoters and their associated enhancer regions (inter-sequence TFBS cooperations).

**Terminology** For the sake of simplicity and to avoid misunderstandings, I adopt the terminology of our papers [43] and [44]. Thereby, a match of a position weight matrix (PWM) with a segment of genomic DNA is termed (potential) transcription factor binding site (TFBS). TFBSs are represented by names of their corresponding PWM. A TFBS pair in the context of intra-sequence cooperating TFs refers to co-occurring TFBSs while a TFBS pair in the context of inter sequence cooperating TFs refers to associated TFBS distributions. In both cases, I can not make any statement about the kind of interaction (cooperativity, synergistic or antagonistic interaction etc.) of the underlying TFs. The term cooperation refers to any kind of functional cooperation and/or physical interactions between the constituents of the predicted TFBS pairs.

### 4.1. Identification of intra-regional cooperating TFs using pointwise mutual information

In higher organisms, the interplay between transcription factors is much more important than the single factor itself. Cooperating TFs tend to bind close to each other on DNA in order to fulfill their regulatory functions. Therefore, the TFBS distribution on DNA provides information about the preferred cooperation partners of single factors. Following this, I developed a method and present it in the following chapter that identifies pairs of TFBSs that significantly co-occur in a set of given sequences. The method is twofold in a way that I first present the general method and afterwards, I extend it in order to identify only sequence set specific TF cooperations. This section is mainly based on our recently published papers [43, 44] (see Appendix A.1 and Appendix A.2).

### 4.1.1. Cooperating TFs

In this section, I introduce the idea for using pointwise mutual information for the identification of cooperating transcription factors based on the co-occurrence in a set of sequences.

**Pre-processing work** In the first step, I obtain all promoter regions for the set of RefSeq genes under study based on their annotated transcription start site (TSS) using UCSC Table Browser [45]. Thereby, I use the hg19 release of the human genome and consider only chromosome annotations of chromosome chr1-chr22, chrX and chrY. Due to the fact that alternative promoter regions for the same gene tend to overlap resulting from the underlying RefSeq annotations, I filter redundant promoters based on their TSS by randomly picking one of the redundant promoter sequences and, consequently, regard only sequences in the analysis which have no overlap.

Afterwards, I predict all potential transcription factor binding sites (TFBSs) in the obtained sequences and their reverse complement using the Match<sup>TM</sup> program by setting the profile parameters as specified by [4]. I further use the PWM library proposed by [4] of TRANSFAC<sup>®</sup> release 2014.1.

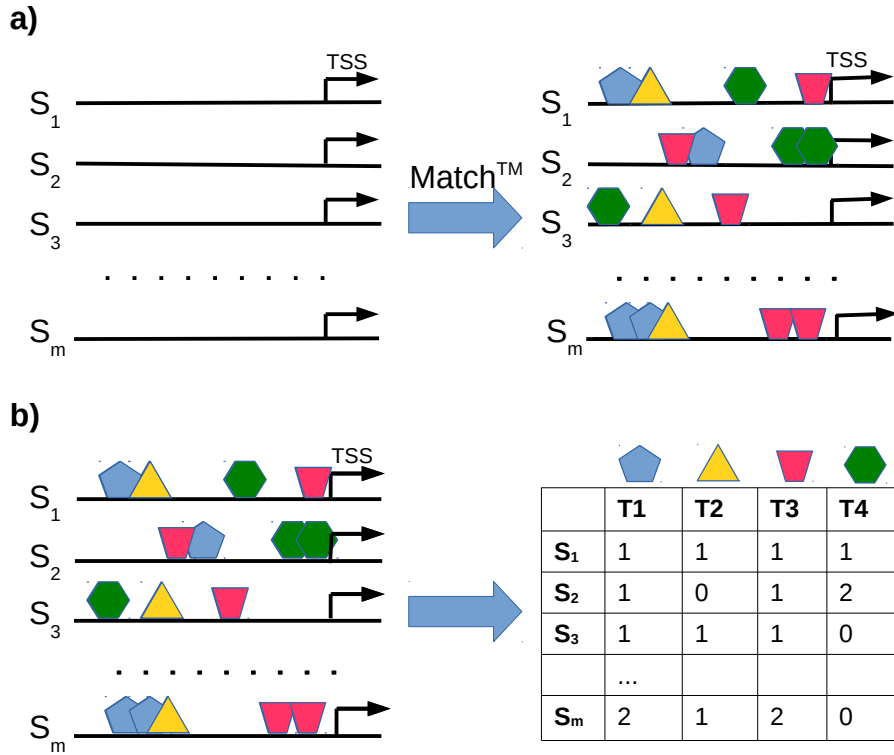
**Workflow** The algorithm for the determination of co-occurring TFBSs comprises six phases that are explained in detail in the following.

**Phase 1: Construction and filtering of TFBS-sequence matrix** Based on the number of predicted TFBSs in each sequence under study a TFBS-sequence matrix  $\mathbb{M}$  is generated where rows correspond to the sequence IDs and columns to the names of PWMs. Thereby, an entry in  $\mathbb{M}$  is defined as follows: Let TFBS  $t_j$  be a TFBS predicted by PWM  $j$  ( $j \in 1, \dots, n$ , where  $n$  is the number of PWMs in the library) and  $s_i$  ( $i \in 1, \dots, m$ , where  $m$  is the number of sequences under study) be a promoter sequence, an entry  $f_{ij}$  in  $\mathbb{M}$  corresponds to the frequency of  $t_j$  in  $s_i$  (see Figure 4.1). It turned out that some TFBSs are highly over-represented, while some other TFBSs occur rarely in a minority of the sequences. In order to reduce the bias of highly represented TFBSs or noisy effects arising from insufficient data the corresponding columns are filtered by removing all columns that i) contain more zero entries than average and ii) having a column sum  $\leq 3 \times \sigma$ , where  $\sigma$  is the standard deviation of all column sums in  $\mathbb{M}$ .

**Phase 2: Identification of important TFBSs in each sequence** Following the idea of linguistics for document summarizing processes, I characterize the important TFBSs for each sequence based on the filtered  $\mathbb{M}$  by calculating the pointwise mutual information ( $\text{PMI}_{s,t}$ ) between a sequence  $s_i$  and a TFBS  $t_j$  as

$$\text{PMI}(s_i, t_j) = \log_2 \frac{p(s_i, t_j)}{p(s_i)p(t_j)}, \quad (4.1.1)$$





**Figure 4.1.: Construction of TFBS-sequence matrix. a)** In a first step, for all sequences under study, all potential transcription factor binding sites are predicted using Match<sup>TM</sup> program [35]. **b)** In the next step, the TFBS-sequence matrix is generated where rows correspond to the promoter sequences and columns to PWMs used for TFBS prediction. An entry in the matrix refers to the frequency of predicted TFBSs in the corresponding sequence. For example, for PWM<sub>4</sub> one corresponding TFBS in sequence  $s_1$  and two TFBSs in  $s_2$  are identified.

where  $p(s_i, t_j)$  is the joint probability for TFBS  $t_j$  occurring in sequence  $s_i$  with respect to the entire sequence set. It is defined as follows:

$$p(s_i, t_j) = \frac{f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (4.1.2)$$

$p(s_i)$  and  $p(t_j)$  are the marginal probabilities of  $s_i$  and  $t_j$ , respectively. They are defined as:

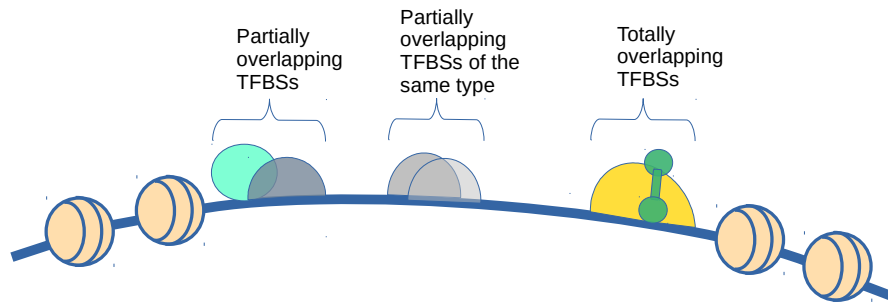
$$p(s_i) = \frac{\sum_{j=1}^n f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (4.1.3)$$

and

$$p(t_j) = \frac{\sum_{i=1}^m f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (4.1.4)$$

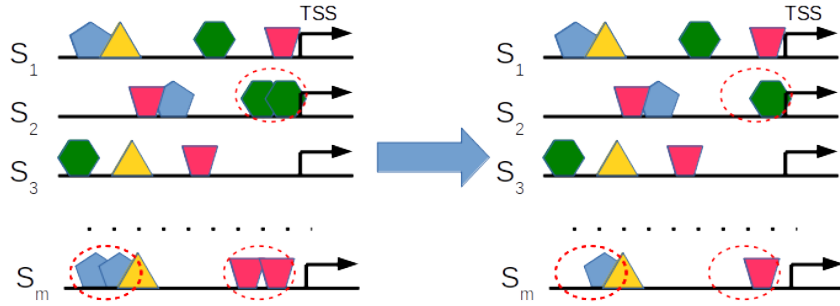
Finally, a TFBS  $t_j$  is considered to be important for sequence  $s_i$  if  $\text{PMII}(s_i, t_j) > 0$  indicating that  $t_j$  occurs more often than expected by pure chance in  $s_i$ . In the following analysis steps, only those TFBSs are considered that have been identified as important by this criterion.

**Phase 3: Filter to avoid overlaps** The Match<sup>TM</sup> algorithm predicts all potential TFBSs based on a PWM library, which can result in multiple predictions for the same sequence region and, thus, overlapping TFBSs. These overlaps can be explained by i) the similarity of some PWMs, ii) the palindromicity of TFBSs (the reverse complement is the same as the original sequence) and iii) some PWMs are larger than the real binding sites of TFs. The overlap of two TFBSs can be partially or a TFBS can totally be included in another binding site (see Figure 4.2). In analogy to [46], I define two TFBSs to be overlapping if their overlapping region exceeds a length of 4 bp.



**Figure 4.2.: Different scenarios for overlapping TFBSs.** On the left side, the binding sites of the blue and the gray transcription factors share a few overlapping nucleotides, while the two gray TFBSs have a large overlapping region. On the right, the binding site of the green TF is totally included in the binding site of the yellow one, indicating that the binding of the two TFs is mutually exclusive.

The overlap of TFBSs of the same type can result in their over-representation in the following analysis steps. Thus, overlapping TFBSs of the same type are filtered in a way that the TFBS survives that has a closer distance to transcription start site (TSS), since the functional important TFBSs are closer to TSS [47]. This filtering process is depicted in Figure 4.3.



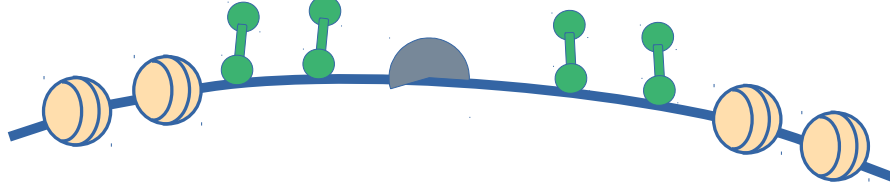
**Figure 4.3.: Filter to avoid overlaps.** Overlapping TFBSs of the same type (marked by dashed circles) are filtered in a way that the TFBS survives that has a closer distance to the transcription start site (TSS) in order to avoid the overestimation of a certain TFBS.

**Phase 4: Construction of TFBS pairs** The distance  $d_{t_A, t_B}$  between two TFBSs  $t_A$  and  $t_B$  is defined as the distance of their centers  $C_{t_A}$  and  $C_{t_B}$ :

$$d_{t_A, t_B} = |C_{t_A} - C_{t_B}| \quad (4.1.5)$$

Thereby, the center  $C_{t_A}$  of a TFBS  $t_A$  is defined as  $\lfloor \frac{length_A}{2} \rfloor$  where  $length_A$  indicates the length of  $t_A$  (see Figure 4.4).

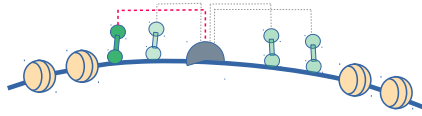
Two TFBSs form a pair if  $d_{min} \leq d_{t_A, t_B} \leq d_{max}$ , where  $d_{min}$  and  $d_{max}$  are pre-defined minimal and maximal distance thresholds, thereby a slight overlap of the TFBSs of at most 4 bps is allowed as suggested in [46]. In this thesis, I set  $d_{min} = 5$  bp which is about half of the length of an average TFBS and tested several different  $d_{max}$  constraints. In the analysis, I have to deal with homotypic clusters, an accumulation of TFBSs of the same type in a certain DNA region that are not necessarily overlapping. This accumulation of a certain TFBS results in a multitude of false positive pairs containing this TFBS. In order to avoid such over-estimations, a TFBS instance can only participate in one pairing of a specified TFBS pair (see the example above for details).

**Example: Homotypic cluster problem**


The green TFBSs  $t_{green}$  form a homotypic cluster and the gray TFBS  $t_{gray}$  is included in the cluster. By simply counting all possible pairs of  $t_{green}-t_{gray}$  result in four pair instances.

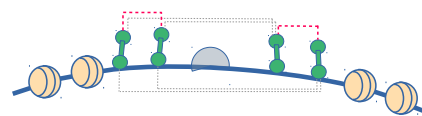
A homotypic cluster is the accumulation of TFBSs of the same type in a certain DNA region. In the example above, the green TFBSs  $t_{green}$  build an homotypic cluster and the gray TFBS  $t_{gray}$  is incorporated in this cluster which leads to four pair instances of  $t_{green}-t_{gray}$ . However, the number of these pairings is an overestimation of the considered pair, since: i) the green binding sites are not all occupied by TFs at the same time and ii) the gray TF can not interact with all green TFs at the same time. In order to avoid this overestimation of TFBS pairs, the pair instances were identified in a way that I consider a certain pair of TFBSs  $t_A$  and  $t_B$  and scan the DNA in 5' - 3' direction to detect instances of this pair. After a certain TFBS  $t_A$  or  $t_B$  is incorporated in a pair instance, it is blocked for additional pairings and cannot participate in another pair instance. Applying this strategy on the example above for the pair  $t_{green}-t_{gray}$  results in one pair instance instead of four.

Considered pair:  $t_{green}-t_{gray}$



Scanning the sequence from left to right for instances of the pair  $t_{green}-t_{gray}$ , the first green TFBS is paired to the gray one (depicted by red lines). Afterwards, the gray TFBS is blocked for additional pairings resulting in just one pair instance of  $t_{green}-t_{gray}$  instead of four (depicted by gray lines).

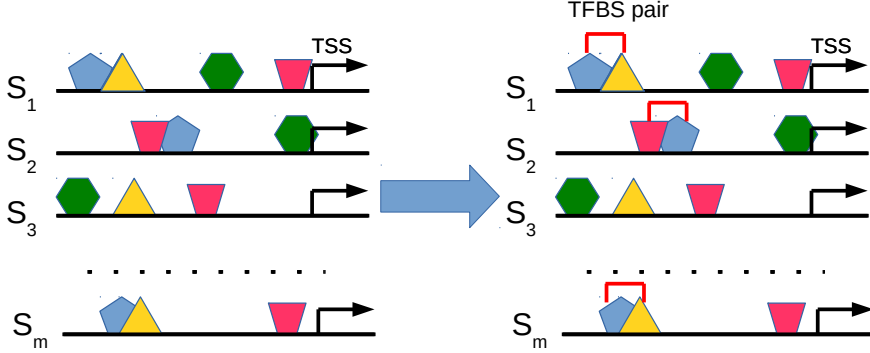
Considered pair:  $t_{green}-t_{green}$



Scanning the sequence from left to right for instances of the pair  $t_{green}-t_{green}$ , the first and the second green TFBSs are paired and afterwards blocked for additional pairs. However, the third and the fourth green TFBSs are not blocked yet and can form an additional pair which results in two pair instances of  $t_{green}-t_{green}$ .

**Phase 5: Weighted cumulative pointwise mutual information** For the identification of potentially collaborating TF pairs, the PMII between all TFBSs  $t_A$  and  $t_B$  is calculated as follows

$$\text{PMII}(t_A; t_B) = \log_2 \frac{p(t_A, t_B)}{p(t_A)p(t_B)} \quad (4.1.6)$$



**Figure 4.4.: TFBS pair construction.** The TFBS pairs were identified based on the distance of their centers and are marked by red lines.

where  $p(t_A; t_B)$  is the joint probability for TFBSs  $t_A$  and  $t_B$  and  $p(t_A)$  and  $p(t_B)$  are the marginal probabilities, respectively. The  $\mathbb{PMII}$  in general is rather susceptible to low number counts [41]. In order to overcome this property to some extent, the  $\mathbb{PMII}(t_A; t_B)$  is scaled by the joint probability  $p(t_A, t_B)$  and the weight  $w_s$  of the corresponding sequence  $s$ , resulting in the weighted pointwise mutual information  $\mathbb{PMII}_p^s(t_A; t_B)$ .

$$\mathbb{PMII}_p^s(t_A; t_B) = w_s \cdot p(t_A, t_B) \cdot \mathbb{PMII}(t_A; t_B) \quad (4.1.7)$$

The weight  $w_s$  of a sequence  $s$  is defined as all TFBS pairs  $N_s$  in  $s$  divided by the total number of TFBS pairs in the sequence set  $S$ .

$$w_s = \frac{N_s}{\sum_{s_i \in S} N_{s_i}} \quad (4.1.8)$$

Finally, in order to determine the important pairs in  $S$ , the  $\mathbb{PMII}_p^s(t_A; t_B)$  for each TFBS pair  $t_A$  and  $t_B$  is summed up over all sequences resulting in the cumulative pointwise mutual information  $\mathbb{PMII}_{pc}(t_A; t_B)$ .

$$\mathbb{PMII}_{pc}(t_A; t_B) = \sum_{s \in S} \mathbb{PMII}_p^s(t_A; t_B) \quad (4.1.9)$$

#### **Phase 6: Background noise reduction of TFBSs using average product correction**

To reduce the effect of false positive TFBS pairs I apply the average product correction (APC) procedure [48] on the  $\mathbb{PMII}_{pc}(t_A; t_B)$  values. I estimate for each TFBS pair  $t_A$  and  $t_B$

the background noise  $APC(t_A, t_B)$  as follows:

$$APC(t_A, t_B) = \frac{\mathbb{PMII}_{pc}(t_A; \bar{t}_x) \cdot \mathbb{PMII}_{pc}(t_B; \bar{t}_x)}{\overline{\mathbb{PMII}_{pc}}}, \quad (4.1.10)$$

where  $\mathbb{PMII}_{pc}(t_A; \bar{t}_x)$  is the average  $\mathbb{PMII}_{pc}$  value of  $t_A$  with all other binding sites and  $\overline{\mathbb{PMII}_{pc}}$  is the overall mean of all calculated  $\mathbb{PMII}_{pc}$  values.  $\mathbb{PMII}_{pc}(t_A; \bar{t}_x)$  is calculated as:

$$\mathbb{PMII}_{pc}(t_A; \bar{t}_x) = \frac{1}{n-1} \sum_{x=1}^n \mathbb{PMII}_{pc}(t_A; t_x), \quad (4.1.11)$$

where  $x = 1, \dots, n$  and  $x \neq a$ . This estimated background noise  $APC(t_A, t_B)$  is then subtracted from the original  $\mathbb{PMII}_{pc}(t_A, t_B)$ -value, resulting in the final  $\mathbb{PMII}_{pc}^{APC}(t_A, t_B)$ -value.

$$\mathbb{PMII}_{pc}^{APC}(t_A, t_B) = \mathbb{PMII}_{pc}(t_A, t_B) - APC(t_A, t_B) \quad (4.1.12)$$

Based on the final  $\mathbb{PMII}_{pc}^{APC}(t_A, t_B)$ -values, the  $z$ -score for each TFBS pair  $t_A$  and  $t_B$  is calculated and a pair is considered as significant, if its  $z$ -score( $t_A, t_B$ )  $\geq 3$ .

The  $z$ -score is calculated as:

$$z\text{-score}(t_A, t_B) = \frac{\mathbb{PMII}_{pc}^{APC}(t_A, t_B) - \overline{\mathbb{PMII}_{pc}^{APC}}}{\sigma_{\mathbb{PMII}_{pc}^{APC}}}, \quad (4.1.13)$$

where  $\overline{\mathbb{PMII}_{pc}^{APC}}$  is the overall mean of  $\mathbb{PMII}_{pc}^{APC}$ -values and  $\sigma_{\mathbb{PMII}_{pc}^{APC}}$  is the corresponding standard deviation.

#### 4.1.2. Sequence-set specific cooperating TFs

In the previous section, I presented a method for the identification of cooperating TFs based on the co-occurrence of their TFBSs using pointwise mutual information (PMII). There, I applied the average product corrections (APC) [48] in order to eliminate the background co-occurrences resulting from false putative TFBS predictions with respect to the entire set of sequences under study. Although the APC appears to be rather successful for its purpose it cannot handle background co-occurrences that stem from common regulatory programs between cell types and common environmental components like GC-content or general (oligo-) nucleotide composition. In order to address this point, I extended the method and estimated the level of background co-occurrences for each TFBS pair by creating a background sequence set that preserves the (oligo-) nucleotide composition of the sequences under study. Thereby, TF cooperations that are sensitive regarding the distance of their TFBSs as well as the specific nucleotide structure have small background values while ubiquitously occurring TF pairs will have larger values, since they are less susceptible to changes in their

binding site distribution. Finally, I will remove the estimated background co-occurrences and thereby separate sequence set specific pairs from common (general important) ones.

**Separation of sequence set specific TF cooperations from the common ones** The overall workflow of the extension approach is depicted in Figure 4.5. In the first step, a sufficiently large number of background sequence sets (e.g. 1000) is created in order to determine the background association of TFBS pairs, where the background sequence sets are constructed in a way that the general nucleotide composition of the sequences and the general assembly of the sequence set (number of sequences and sequence length) is maintained. In this respect, I create a background set by shuffling each sequence of the set under study using the uShuffle algorithm [36] and preserving the number of tri-nucleotides and, thus, the core of TFBSs by setting  $k = 3$ .

Afterwards, all potential TFBSs in the background sequences are predicted using the Match<sup>TM</sup> algorithm and conduct phase one to six of the original approach (see Section 4.1.1 for details):

- **Phase 1:** Construction and filtering of TFBS-sequence matrix
- **Phase 2:** Identification of important TFBSs in each sequence
- **Phase 3:** Filter to avoid overlaps
- **Phase 4:** Construction of TFBS pairs
- **Phase 5:** Weighted cumulative pointwise mutual information
- **Phase 6:** Background noise reduction of TFBSs using average product correction

This results in a  $\mathbb{PMII}_{pc}(t_A; t_B)$ -value for each TFBS pair  $t_A$  and  $t_B$  for each background sequence set. In the next step, the average  $AVG(\mathbb{PMII}_{pc}(t_a; t_b))$  values for two TFBSs  $t_a$  and  $t_b$  are calculated over all background sequence sets:

$$AVG(\mathbb{PMII}(t_a; t_b)) = \frac{1}{l} \sum_{i=1}^l \mathbb{PMII}_{pc}^{APC}(t_a; t_b)_i, \quad (4.1.14)$$

where  $l$  is the number of background sequence sets. After that the  $AVG(\mathbb{PMII}_{pc}(t_a; t_b))$  value of a TFBS pair  $t_a$  and  $t_b$  is subtracted from the original  $\mathbb{PMII}_{pc}(t_a; t_b)$  value of the sequence set under study, resulting in  $\mathbb{PMII}^{specific}(t_a; t_b)$ .

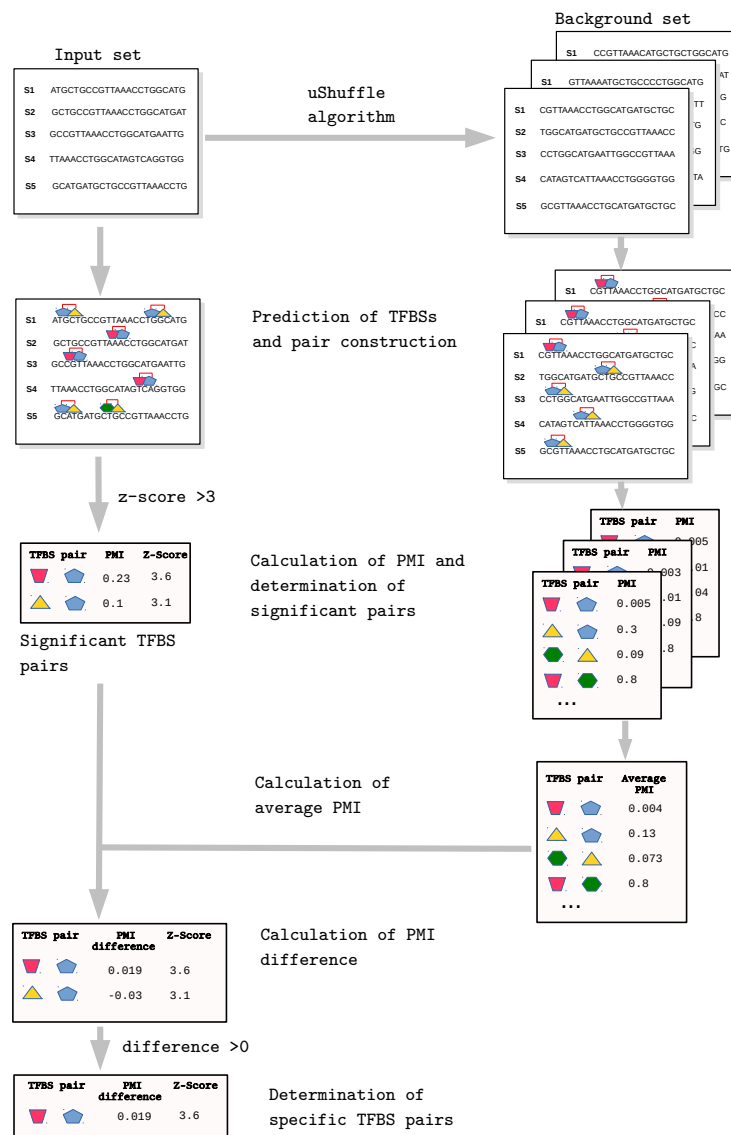
$$\mathbb{PMII}^{specific}(t_a; t_b) = \mathbb{PMII}_{pc}^{APC}(t_a; t_b) - [(1 + \alpha) \cdot AVG(\mathbb{PMII}(t_a; t_b))] \quad (4.1.15)$$

The scaling parameter  $\alpha \in [-1, +1]$  is preassigned and used to enlarge or reduce the influence of the subtracted background level. Thereby, an  $\alpha$  value of -1 results in the original predictions of the original approach while setting  $\alpha = 1$  is the subtraction of the  $AVG(\mathbb{PMII}_{pc}(t_a; t_b))$ -values from the original ones. Setting  $\alpha > 0$  increases the effect of the subtracted background level and leads to a more strict selection process of specific TFBS

pairs. However, in order to avoid the overestimation of the calculated background, the determination of an upper bound for  $\alpha$  is essential. By carefully analyzing the influence of  $\alpha$ , I conclude to set  $\alpha = 1$  as an appropriate upper bound.

Finally, a TFBS pair  $t_a$  and  $t_b$  is defined to be specific for a set of sequences if it is significant according to the original approach ( $z\text{-score} \geq 3$ ) and if  $\mathbb{PMI}^{specific}(t_a; t_b) > 0$ . Significant pairs that have a  $\mathbb{PMI}^{specific}(t_a; t_b) \leq 0$  are considered as general important TFBS co-occurrences.





**Figure 4.5.: Workflow of the extension approach for the determination of sequence set specific TFBS pairs.** In the first step, a sufficiently large number of background sets is created by shuffling the input set. Afterwards, the original method for the determination of significant co-occurring TFBSs is applied and the PMI between all TFBSs is calculated for input and background sets. In the next step, the average PMI for the TFBS pairs of the background sets are determined and afterwards, the background is subtracted from the original PMI-values of the input set. Finally, all significant TFBS pairs that have a positive PMI-difference are defined to be sequence set specific.



## 4.2. Identification of inter-regional associated TFs using multivariate mutual information

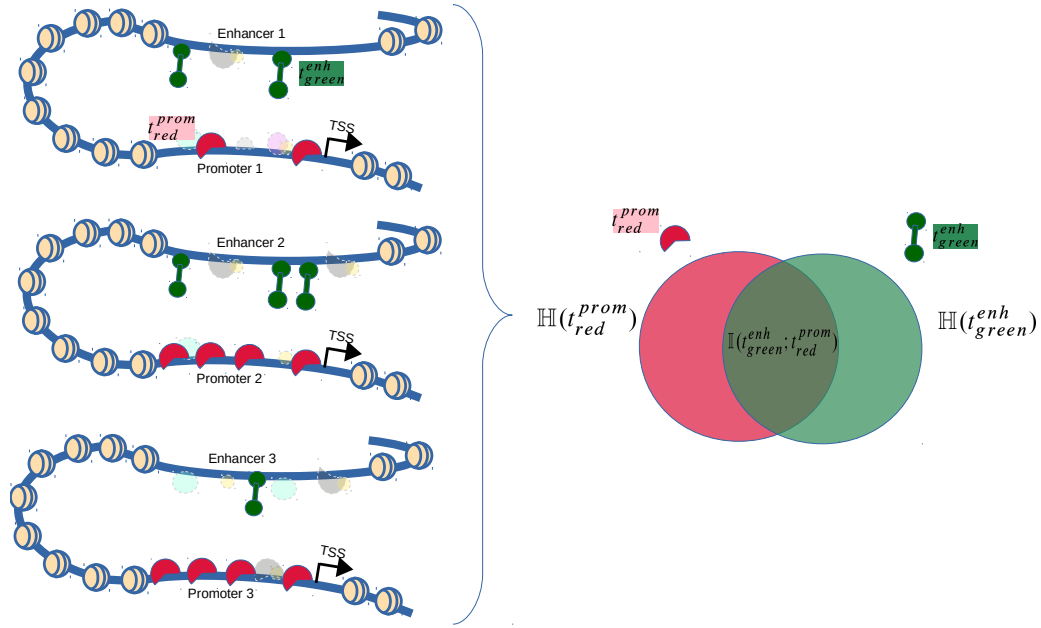
Since the cooperation of transcription factors plays a fundamental role in the establishment of enhancer-promoter interactions (PEIs), the identification of the pairwise association of factors between enhancer and promoter regions is essential for the understanding of gene regulation. To this aim, I utilize the property that the underlying binding site distributions of these factor pairs show a certain degree of association to each other, which can in turn be used for the identification of these associated transcription factors. In this section, I present a method for the identification of associated transcription factors between enhancer and their related promoter regions using an extended form of pairwise mutual information based on predicted transcription factor binding sites (TFBSs) in the sequences of known PEIs (e.g. ChIA-PET data) (see Figure 4.6).

The overall workflow comprises five phases that are explained in detail below. In short: first of all a background set is generated by shuffling the input sequences and all potential TFBSs are predicted in input and background sequences using Match<sup>TM</sup>-algorithm. Based on these predictions, a TFBS-sequence count matrix is generated for enhancer and promoter sequences, respectively, where the count values are normalized and assigned to intervals in the next step. Based on these intervals, different mutual information quantities are used in order to determine associated TFBS pairs between enhancer and promoter regions.

**Phase 1: Creation of background sequences** Given a set of known PEIs, each sequence (of promoter and enhancer region) is shuffled using the uShuffle [36] algorithm and set  $k = 3$ . Thereby, the number of tri-nucleotides and, thus, the core of TFBSs as well as the sequence length is maintained. The resulting pairs of shuffled enhancer and promoter sequences are used as background pairings in the following steps. In order to differentiate between input and background sequences, I further define a vector  $\mathbb{V}^{label}$  that contains the origin of the underlying PEI ( $I$  for an input PEI and  $B$  for background sequences).

**Phase 2: Determination of TFBS-sequence count matrices** In order to identify potential transcription factor binding sites (TFBSs) in the sequences under study as well as in the shuffled sequences, the Match<sup>TM</sup> algorithm [35] is applied using the *minimizing the number of false positive predictions* (minFP) profile.

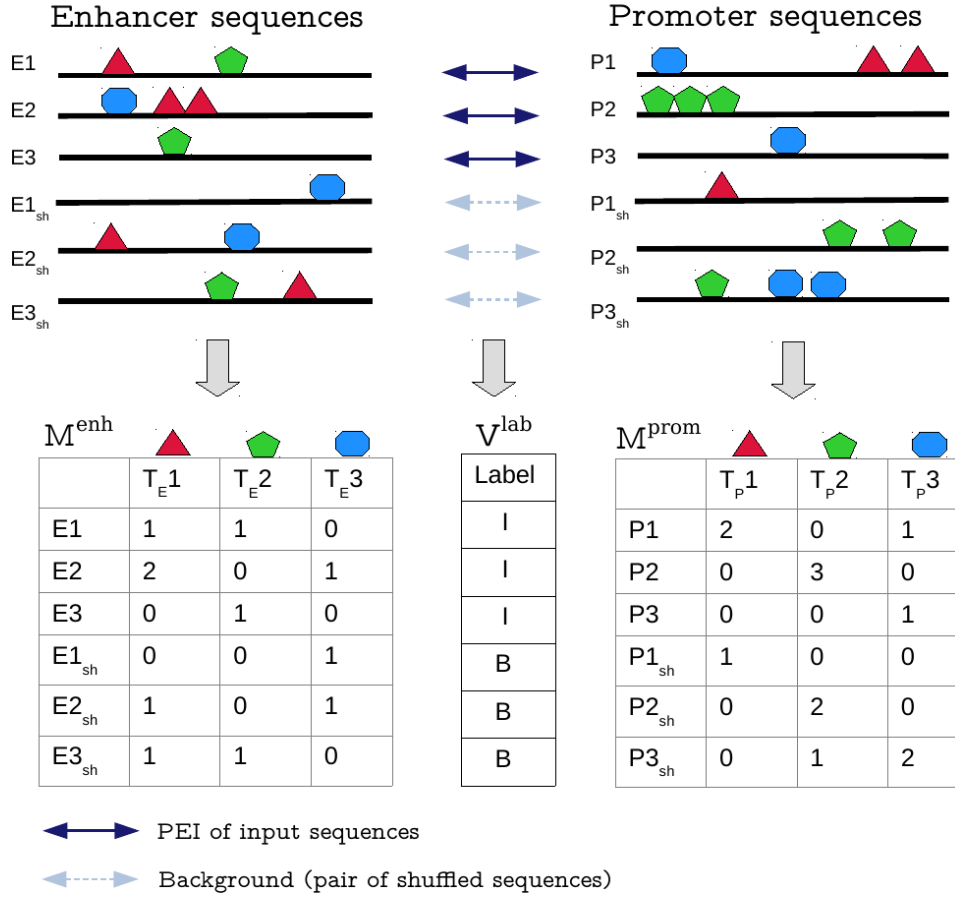
In analogy to Chapter 4.1, I created TFBS-sequence matrices  $\mathbb{M}^{enh}$  and  $\mathbb{M}^{prom}$  for enhancer and promoter sequences, respectively. In both matrices the columns correspond to the position weight matrices (PWMs) and rows to sequences of enhancer and promoter, respectively. An entry  $f_{ij}$  in  $\mathbb{M}$  is the frequency of TFBS  $t_j$  predicted by PWM  $j$  ( $j \in 1, \dots, m$ , where  $m$  is the number of PWMs under study) in sequence  $s_i$  ( $i \in 1, \dots, n$ , where  $n$  is the number of sequences under study). The rows of the matrices corresponding to enhancer and promoter



**Figure 4.6.: Identification of associated TFs between enhancer and promoter sequences using mutual information  $\mathbb{I}$ .** On the basis of a set of paired enhancer promoter sequences (left), all potential TFBSs were predicted and the mutual information between a certain pair of TFBSs between enhancer and promoter region (i.e.  $t_{green}^{enh}$  and  $t_{red}^{prom}$ ) is calculated (right) based on the frequency of occurrences in the underlying sequences.

sequences are ordered as follows: row  $i$  of both,  $\mathbb{M}^{enh}$  and  $\mathbb{M}^{prom}$ , corresponds to the index of the corresponding PEI  $i$  (see Figure 4.7).

**Phase 3: Normalization and interval building** The overall aim is to calculate information theoretic quantities between two TFBS distributions and consequently, probability mass distributions are required based on an alphabet that reflects the count values for the TFBSs in the sequences. However, it is not possible to use the count values itself as letters in the alphabet, since there are too many count values and the separation between count values that differ by one is not appropriate. Therefore, I decided to assign each count value to an interval and use the interval identifiers  $z^k$  (for  $k = 1, \dots, q+1$  and  $z \in \mathfrak{Z}$ ,  $\mathfrak{Z} = \{-1, 1, 2, \dots, q\}$ ) as letters in the alphabet. In the first step of this process, I construct  $q+1$  intervals where  $q$  intervals are in  $[0, 1]$  as  $((0, \frac{1}{q}], (\frac{1}{q}, \frac{2}{q}], \dots, (\frac{q-1}{q}, 1])$  and one is used for zero count values. In the next step, the count values  $f_{ij}$  are normalized by *min-max normalization* in order to scale them in the range between zero. As generally known, the *min-max normalization* is



**Figure 4.7.: Determination of TFBS-sequence count matrices.** All potential TFBSs were predicted in each sequence (input sequences E1-3 and P1-3 as well as shuffled sequences E1-3<sub>sh</sub> and P1-3<sub>sh</sub>) using Match<sup>TM</sup>-algorithm. Afterwards, for enhancer and promoter sequences, TFBS-sequence count matrices  $M^{enh}$  and  $M^{prom}$  are created, respectively, where a matrix entry refers to the frequency of predicted binding sites in the corresponding sequence. The vector  $V^{lab}$  is important to differentiate between input and background sequences.

defined as

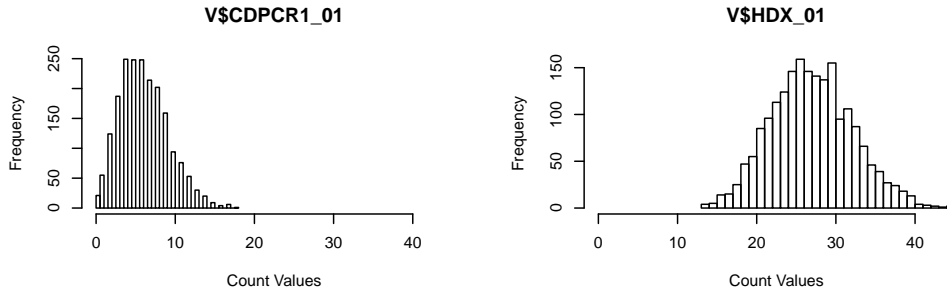
$$f_{ij}^{norm} = \frac{f_{ij} - f_{min}}{f_{max} - f_{min}}, \quad (4.2.1)$$

where  $f_{min}$  and  $f_{max}$  are the minimum and maximal values of  $\mathbb{M}$ , respectively. Finally, I assign each normalized count value  $f_{ij}^{norm}$  into the appropriate interval  $z_{ij}$ , following:

$$z_{ij} = \begin{cases} \lfloor f_{ij}^{norm} \times q \rfloor & \text{if } f_{ij}^{norm} > 0 \\ -1, & \text{otherwise,} \end{cases} \quad (4.2.2)$$

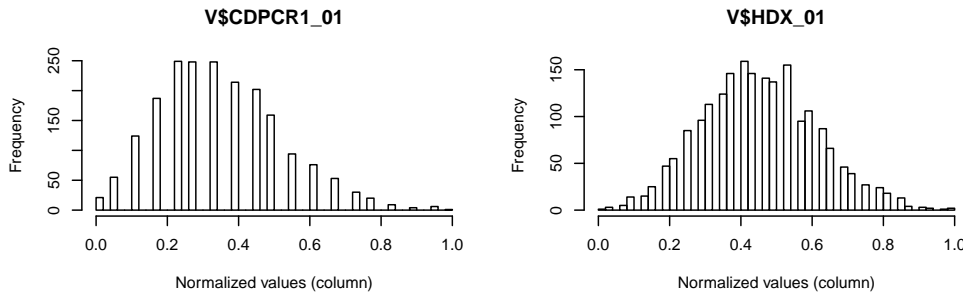
To this end, I define the interval-sequence matrices  $\mathbb{M}_{int}^{enh}$  and  $\mathbb{M}_{int}^{prom}$  where an entry  $z_{ij}$  in  $\mathbb{M}_{int}$  is the interval identifier of which the count value  $f_{ij}^{norm}$  of TFBS  $t_j$  in sequence  $s_i$  is assigned to. The transformation process from TFBS-sequence matrix  $\mathbb{M}$  to interval-sequence matrix  $\mathbb{M}_{int}$  is depicted in Figure 4.8.

### Example: Influence of normalization strategies on count value distribution



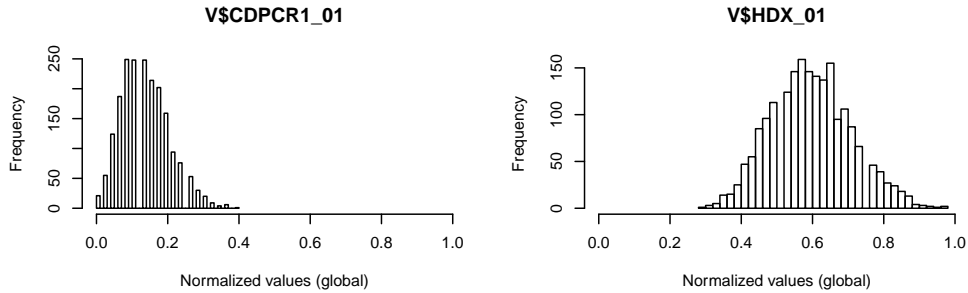
Distribution of count values (frequency of TFBSs per sequence) for V\$CDPCR1\_01 and V\$HDX\_01.

In this example, I show the influence of the global and column-wise *min-max normalization* strategies on the count value distributions of TFBSs V\$CDPCR1\_01 and V\$HDX\_01. Regarding the original count values, both distributions resemble a Poisson distribution that differ in mean and variance. Normalizing the count values of each TFBS on its own (column wise) results for both TFBSs in Poisson distributions in the range between 0 and 1 that have a similar mean and variance. Consequently, the information of variance and value range of the original count value distributions disappears and a differentiation between the two distributions is hard.



*Distribution of column wise normalized count values for V\$CDPCR1\_01 and V\$HDX\_01.*

In the global *min-max normalization* strategy, the global minimum and maximum value of the entire count matrix is used resulting in normalized values in the range between 0 and 1 where the proportions between mean and variance inside a distribution as well as between different distributions are maintained.

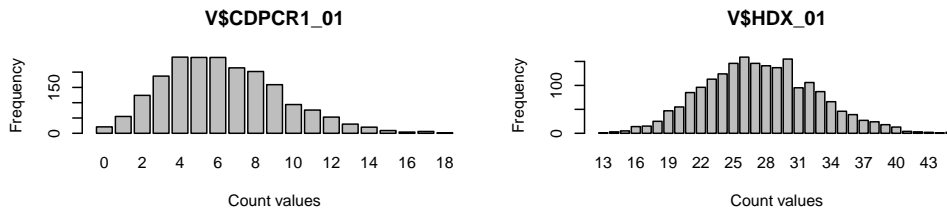


*Distribution of global normalized count values for V\$CDPCR1\_01 and V\$HDX\_01.*

To this end, the global normalization strategy enables a differentiation of the normalized count value distributions of different TFBSs regarding the spread of the data as well as their general count value proportions in comparison to other TFBSs.

In order to normalize the count values by maintaining the information of value range and dispersion from the mean, I decided to use the global *min-max normalization* strategy in the analysis.

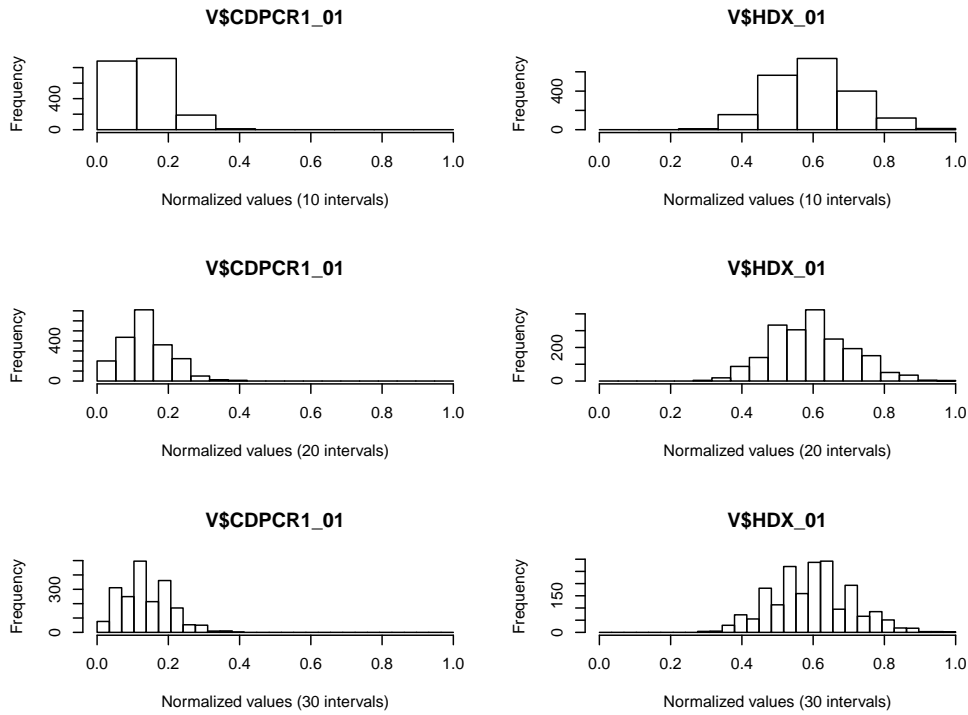
**Example: Influence of interval size on normalized count value distribution**



*Distribution of count values (freq. of TFBSs/sequence) for V\$CDPCR1\_01 and V\$HDX\_01.*

In the interval building process, the range between 0 and 1 is divided into  $q$  intervals of the same size where the normalized values  $f_{ij}^{norm}$  are assigned into. The interval identifiers  $z^k$  (for  $k = 1, \dots, q + 1$ ) form then the letters of the alphabet  $\mathfrak{Z} = \{-1, 1, 2, \dots, q\}$ . Choosing an alphabet size of  $q = 10$ , the range between 0 and 1 is divided into ten intervals of the same size (e.g. 0.1). By increasing the number of intervals, the size of the intervals itself is decreasing. A small number of intervals implicates an accumulation of normalized values in a few intervals. In example, by choosing  $q = 10$ , for V\$CDPCR1\_01 four intervals are occupied by normalized values while the remaining six intervals are empty. Further, the majority of values is assigned to two intervals, while there are only a few values assigned to the two other intervals. Con-

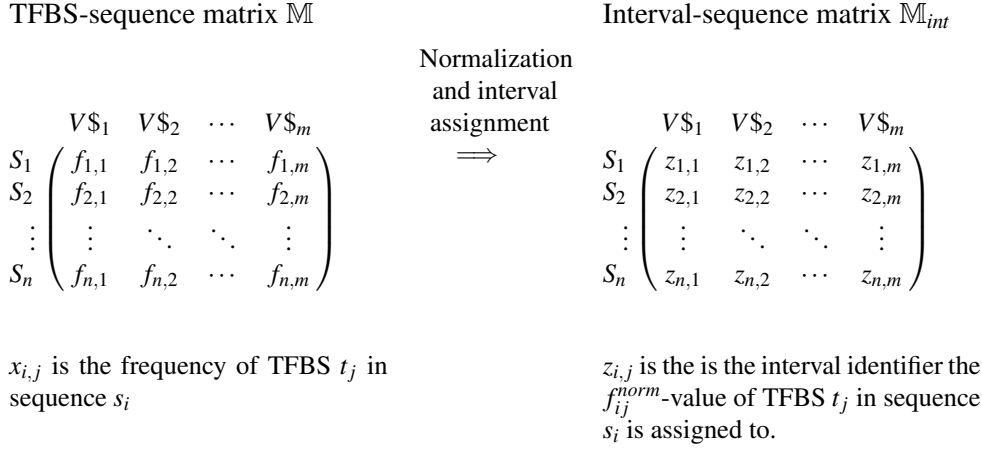
sequently, by grouping the values using only a few intervals, the differentiation between the individual  $f_{ij}^{norm}$  values as well as their frequency distribution gets blurred. By increasing the number of intervals, the differentiation between the values becomes more clear. However, a huge number of intervals leads to an overestimation of small differences in the count values.



*Distribution of normalized count values assigned to 10, 20 and 30 intervals (depicted as bars) that are equally distributed in the range between 0 and 1 for V\$CDPCR1\_01 and V\$HDX\_01.*

The overall aim of the normalization and interval building process was to drop down the alphabet size of each TFBS to a more discrete level and at the same time to maintain the differentiation between count value distributions. The binding site V\$CDPCR1\_01 takes 18 different count values, for binding site V\$HDX\_01 there are 33 different count values. After normalizing the values and assigning each of them to ten intervals equally distributed in  $[0, 1]$  the values of V\$CDPCR1\_01 take four intervals, while those of V\$HDX\_01 are assigned to 7 intervals. Increasing the number of intervals to 20 and 30, the values of V\$CDPCR1\_01 took 8 and 11 intervals, respectively, while those of V\$HDX\_01 took 14 and 21 intervals. Thus, choosing an interval size of 30 leads to a reduction of the alphabet size of about 2/3 of the original size.





**Figure 4.8.: Conversion of TFBS-sequence matrix to interval-sequence matrix.** The TFBS-sequence matrix  $\mathbb{M}$  captures the count values  $f_{ij}$  of a TFBS  $t_j$  in a sequence  $s_i$ . In Phase 3, each count value is normalized and afterwards assigned into the appropriate interval resulting in an interval-sequence matrix  $\mathbb{M}_{int}$  that captures for each TFBS  $t_j$  in sequence  $s_i$  the interval identifier  $z_{ij}$  of the assigned interval.

**Phase 4: Construction of interval matrix and probability mass functions** The empirical distribution  $p_{t_j}(z^k)$  of interval  $z^k$  based on  $\mathbb{M}_{int}$  for a TFBS  $t_j$  is defined as

$$p_{t_j}(z^k) = \frac{\#(z^k)}{n} \quad (4.2.3)$$

where  $\#(z^k)$  is the frequency of interval  $z^k$  observed for TFBS  $t_j$  and  $n$  is the number of sequences. The joint probability of intervals  $z^k$  and  $z^l$  (for  $k, l = 1, \dots, q+1$ ) of TFBSs  $t_{j_1}^{enh}$  and  $t_{j_2}^{prom}$ , respectively, is defined to be:

$$p_{t_{j_1}^{enh}, t_{j_2}^{prom}}(z^k, z^l) = \frac{\#(z^k, z^l)}{n}, \quad (4.2.4)$$

where  $\#(z^k, z^l)$  is the frequency of joint occurrences of interval  $z^k$  and  $z^l$  in the paired enhancer and promoter sequences.

Additionally, the probabilities of the label vector  $\mathbb{V}^{label}$  are  $p(l) = 0.5$  for  $l \in L$  and  $L = \{B, I\}$ .

**Phase 5: Calculation of mutual information** In order to identify associated TFs I calculate the mutual information between their underlying binding site distributions based on the interval matrices  $\mathbb{M}_{int}^{enh}$  and  $\mathbb{M}_{int}^{prom}$ .

The pairwise mutual information  $\mathbb{I}(t_j^{enh}, t_k^{prom})$  between a TFBSs  $t_{j_1}^{enh}$  predicted in an enhancer and a TFBS  $t_{j_2}^{prom}$  in a promoter sequence is defined as

$$\mathbb{I}(t_{j_1}^{enh}, t_{j_2}^{prom}) = \sum_{k \in \mathfrak{Z}} \sum_{l \in \mathfrak{Z}} p_{t_{j_1} t_{j_2}}(z^k, z^l) \log \frac{p_{t_{j_1} t_{j_2}}(z^k, z^l)}{p_{t_{j_1}}(z^k) p_{t_{j_2}}(z^l)}, \quad (4.2.5)$$

where the joint probability  $p_{t_{j_1} t_{j_2}}(z^k, z^l)$  of two intervals  $z^k$  and  $z^l$  and the marginal probabilities  $p_{t_{j_1}}(z^k)$  and  $p_{t_{j_2}}(z^l)$  are computed like explained above. In the analysis, I have to deal with three random variables, since the differentiation between input and background set is necessary. Thus, I incorporate the label of the origin of the paired sequences (stored in  $\mathbb{V}^{lab}$  as a third variable).

As presented in Chapter 3, there are several different multivariate mutual information quantities that deal with three random variables and I consider all of them in the analysis in order to find the best quantity for the approach. For two TFBSs  $t_{j_1}^{enh}$  and  $t_{j_2}^{prom}$  in enhancer and promoter sequences, respectively and with regard of the label  $L$  that contains the origin of the underlying sequences, the quantities are defined as follows:

- Multivariate mutual information
  - $\text{MMI}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$
- Joint mutual information
  - $\text{JMI}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$
- Conditional mutual information
  - $\text{CMI}(t_{j_1}^{enh}, t_{j_2}^{prom} | L)$
- Dual total correlation
  - $\text{DTC}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$

Afterwards, I normalize the resulting values using the logarithm of the maximal alphabet size ( $\log_2(\max|\mathfrak{X}|, |\mathfrak{Y}|)$ ) in order to provide a better comparison between the results and to eliminate side effects like alphabet size.

### Normalization strategies of mutual information quantities

Case 1			Case 2			Case 3		
$t_{j1}^{enh}$	$t_{j2}^{prom}$	<b>L</b>	$t_{j1}^{enh}$	$t_{j2}^{prom}$	<b>L</b>	$t_{j1}^{enh}$	$t_{j2}^{prom}$	<b>L</b>
1	2	I	1	0	I	1	4	I
1	2	I	1	0	I	1	4	I
1	2	I	2	1	I	1	4	I
1	2	I	2	1	I	2	4	I
3	0	B	3	2	B	3	5	B
3	0	B	3	2	B	3	1	B
3	0	B	4	3	B	3	1	B
3	0	B	4	3	B	3	1	B

*Example for the alphabet size effect. Case 1 and Case 2 show a perfect correlation, although in Case 2 there are more different entities than in Case 1. Case 3 does not show such a perfect association between the columns at all.*

All mutual information quantities depend to some extent on the alphabet size and the entropy of the distributions under study, here, the columns of a matrix. If two columns show a perfect association (as shown in Case 1 and Case 2) the column pair with larger alphabet size has the larger mutual information. This, in turn, can lead to wrong interpretations of the results and can be avoided by normalizing the mutual information values. There are several possibilities to normalize mutual information that are all related to upper thresholds of the measures. The most common normalization strategies are: i) the maximal alphabet normalization ( $\log_2(\max\{|\mathcal{X}|, |\mathcal{Y}|\})$ ), ii) the joint entropy ( $\mathbb{H}(t_{j1}^{enh}, t_{j2}^{prom}, L)$ ) and iii) the sum of the marginal entropies ( $\mathbb{H}(t_{j1}^{enh}) + \mathbb{H}(t_{j2}^{prom}) + \mathbb{H}(L)$ ).

Quantity	Normalization strategy	Case 1	Case 2	Case 3
$\mathbb{I}(t_{j_1}^{enh}; t_{j_2}^{prom})$	-	0	1	0
	$\log_2(\max\{ \mathcal{X} ,  \mathcal{Y} \})$	0	0.5	0
	$\mathbb{H}(t_{j_1}^{enh}, t_{j_2}^{prom})$	0	0.5	0
	$\mathbb{H}(t_{j_1}^{enh}) + \mathbb{H}(t_{j_2}^{prom})$	0	0.2	0
$\mathbb{MMI}(t_{j_1}^{enh}, t_{j_2}^{prom}; L)$	-	1	1	1
	$\log_2(\max\{ \mathcal{X} ,  \mathcal{Y} ,  \mathcal{L} \})$	1	0.5	0.63
	$\mathbb{H}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$	1	0.5	0.55
	$\mathbb{H}(t_{j_1}^{enh}) + \mathbb{H}(t_{j_2}^{prom}) + \mathbb{H}(L)$	0.33	0.2	0.22
$\mathbb{JMI}(t_{j_1}^{enh}, t_{j_2}^{prom}; L)$	-	1	1	1
	$\log_2(\max\{ \mathcal{X} ,  \mathcal{Y} ,  \mathcal{L} \})$	1	0.5	0.63
	$\mathbb{H}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$	1	0.5	0.55
	$\mathbb{H}(t_{j_1}^{enh}) + \mathbb{H}(t_{j_2}^{prom}) + \mathbb{H}(L)$	0.33	0.2	0.22
$\mathbb{CMI}(t_{j_1}^{enh}; t_{j_2}^{prom}   L)$	-	0	1	0
	$\log_2(\max\{ \mathcal{X} ,  \mathcal{Y} ,  \mathcal{L} \})$	0	0.5	0
	$\mathbb{H}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$	0	0.5	0
	$\mathbb{H}(t_{j_1}^{enh}) + \mathbb{H}(t_{j_2}^{prom}) + \mathbb{H}(L)$	0	0.2	0
$\mathbb{DTTC}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$	-	1	2	1
	$\log_2(\max\{ \mathcal{X} ,  \mathcal{Y} ,  \mathcal{L} \})$	1	1	0.63
	$\mathbb{H}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$	1	1	0.55
	$\mathbb{H}(t_{j_1}^{enh}) + \mathbb{H}(t_{j_2}^{prom}) + \mathbb{H}(L)$	0.33	0.4	0.2

*Influence of normalization strategies on the mutual information quantities for the three cases.*

*Shown are the results for the different quantities without normalization (-) and with normalization strategies: maximal alphabet size ( $\log_2(\max\{|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{L}|\})$ ), joint entropy ( $\mathbb{H}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$ ) and the sum of marginal entropies ( $\mathbb{H}(t_{j_1}^{enh}) + \mathbb{H}(t_{j_2}^{prom}) + \mathbb{H}(L)$ ).*

The table shows the results of the different normalization strategies for the different quantities based on the given example. Regarding  $\mathbb{I}(t_{j_1}^{enh}; t_{j_2}^{prom})$  as well as  $\mathbb{CMI}(t_{j_1}^{enh}; t_{j_2}^{prom} | L)$ , the different normalization strategies do not alter the order of the three cases. However, the ratio between the values has been decreased. The  $\mathbb{MMI}(t_{j_1}^{enh}, t_{j_2}^{prom}; L)$  and the  $\mathbb{JMI}(t_{j_1}^{enh}, t_{j_2}^{prom}; L)$  metric results without normalization in the same value for all three cases. Normalizing these values leads to the highest value for Case 1, and the lowest value for Case 3 for all strategies. Thereby, the ratio between the values is high using alphabet normalization strategy and low for using the sum of entropies. For the  $\mathbb{DTTC}(t_{j_1}^{enh}, t_{j_2}^{prom}, L)$  the normalization effects the order of the cases regarding their  $\mathbb{DTTC}$ -value. Without normalization, Case 2 has the highest  $\mathbb{DTTC}$ -value. Using the alphabet size or the joint entropy for normalization, Case 1 and Case 2 are on the same rank. In contrast, using the sum of marginal entropies, again, Case 2 appears to be best associated. In the following, I use the alphabet size  $\log_2(\max\{|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{L}|\})$  for the normalization of all quantities, since all normalization strategies deliver comparable results.

## 5. Results

In this chapter I present the results of the application of my two approaches by first focusing on the identification of potentially intra-sequence cooperating transcription factors and second, on inter-sequence associated transcription factor pairs. For each approach, I demonstrate the performance of the methodology on a simulated dataset and show a comparison to existing methods. Further, I applied the method to real biological data in order to get new insights regarding the gene regulating mechanisms of biological systems and present the results of these biological analyses.

### 5.1. Identification of intra-regional cooperating TFs using pointwise mutual information

In this section, I present the results for the identification of collaborating transcription factors based on the co-occurrence of their binding sites using pointwise mutual information (PMI). In the first part, I present the results of the general approach and in the second part, I demonstrate the performance of the extension approach for the separation sequence-set specific TFBS pairs from the general important ones in a comparative manner to the original (general) approach. This section is mainly based on our recently published papers [43, 44] (see Appendix A.1 and Appendix A.2).

#### 5.1.1. Cooperating TFs

**Data** In order to apply the method to biological data, I analyzed two datasets, a genome wide and a breast cancer gene set. Performing a genome wide analysis, I took all annotated transcription start sites (TSS) of human RefSeq genes and selected the promoter regions 1000bp upstream of TSS and 100bp downstream.

The second dataset is a breast cancer associated gene set that was taken from Joshi et al. [49]. Following [49] I performed the analysis on the promoter regions 500bp upstream and 100bp downstream from the TSS of the corresponding RefSeq genes. For both promoter sets, I made sure that there are no overlapping sequences to avoid their overestimation in the analysis.

**Comparison with existing methods** For a comprehensive evaluation of the method I performed a pairwise comparison with the existing methods: CPModule [7], CrmMiner

[13] and MatrixCatch[4]. I will shortly introduce the methods in the following. For details, please have a look at the original publications.

CPModule was developed by Sun et al. in 2012 and is a method for the detection of unstructured cis-regulatory modules based on *constrained programming for itemset mining framework*. The method requires a set of sequences and a PWM library as input and selects afterwards motifs that i) occur frequently in the input sequences (frequently constraint) ii) are localized within a certain distance on DNA (proximity constraint), iii) are non-redundant (redundancy constraint) and as an optional constraint iv) the module contains a query motif (query-based constraint). All modules that fulfill these constraints are validated by p-values that expresses their specificity for the input set in consideration of the whole genome as background sequence set [7].

CrmMiner has been published by Girgis et al. in 2012 for the determination of enriched motif pairs specific for a given set of sequences. The algorithm requires a mixed and a control set as input where the mixed set contains the regulatory sequences under study and the control set is a set of randomly selected genomic sequences. The input sequences are distributed in a training, a validation and a test set. In the training phase of the algorithm, enriched motifs pairs are identified and the sequences containing these pairs are selected in both, mixed and control set. Afterwards, the sequences are scored according to their enriched motif pairs and based on these scores a Bayesian classifier is trained in order to differentiate between sequences in the mixed and in the control set. In the validation phase parameters are optimized using the validation set. Training and validation phase are repeated until the parameter setting is optimized and the results are later evaluated by using the test set [13].

MatrixCatch was developed in 2013 by Deyneko et al. for the recognition of composite elements provided by TRANSCompe1<sup>®</sup> [14] data base in a given set of sequences. In order to identify these transcription factor pairs in the sequences, the composite elements are modeled by two PWMs, their minimal matching scores, relative orientations and distance constraints [14]. The input sequences are scanned for the identification of composite elements in the sequences and for each composite element match a p-value is calculated in order to determine its recognition quality [4]. Interested readers are kindly referred to our recent study [50] (see Appendix A.4) about the identification of stage-specific transcription factor clusters in human heart development using MatrixCatch in combination with the Markov clustering algorithm.

I compared the performance of my approach with that of the existing tools in two different ways. First of all I generated a simulation dataset by artificially inserting a TFBSs pair, second, I applied all methods to the genome wide and the breast cancer gene set and performed a statistical analysis and a comparison study between the predictions of the different methods. In both cases, I applied all methods using the same position weight matrix (PWM)

**Table 5.1.: Total number of predicted TFBS pairs for the genome wide and the breast cancer analysis** of my approach with maximal distances 20bp, 50bp and 100bp ( $PC-TraFF_{20}$ ,  $PC-TraFF_{50}$  and  $PC-TraFF_{100}$ ), MatrixCatch (MC), CPModule (CPM) and CrmMiner (CrmM).

	Total number of predicted TFBS pairs					
	$PC-TraFF_{20}$	$PC-TraFF_{50}$	$PC-TraFF_{100}$	MC	CPM	CrmM
Genome-wide analysis	54	86	91	19	17	21
Breast cancer analysis	64	82	88	13	6	25

library as suggested in [4], in order to make the results comparable to each other. I ran the comparison tools using their default settings.

For the creation of the simulation dataset, I randomly picked 200 promoter sequences of RefSeq genes of chromosome 21 and inserted the TFBS pair (V\$USF\_01 - V\$IRF1\_01) two to twelve times into these sequences. Thereby, I used the consensus sequences of both motifs and defined the distance between the inserted motifs to be at least 5 bps and at most 20 bps. I applied my approach and the other three methods to this simulation dataset where the inserted pair was successfully predicted by CPModule and my approach, but was not identified by any of the other methods.

Second, I performed a pairwise comparison study of the different approaches by analyzing the breast cancer gene set as well as the genome wide gene set with my approach and the existing methods. In my approach, two TFBSs form a pair if the distance of their centers is between a predefined minimal and maximal distance. In the course of this comparison study, I applied my approach using different maximal distances between the TFBSs (20bp, 50bp and 100 bp) that are indicated as  $PC-TraFF_{20}$ ,  $PC-TraFF_{50}$  and  $PC-TraFF_{100}$  in the following, in order to further evaluate the influence of the maximal distance constraints.

Applying the methods to the breast cancer and the genome wide gene set, it is remarkable that all methods show different numbers of predicted pairs. While my approach identifies a comparably large number of pairs (see Table 5.1) the number of pairs identified by the other approaches is between six and 25 pairs. This phenomenon can be explained by the underlying methodologies. The number of detected pairs by MatrixCatch is restricted to the collection of pairs in the TRANSCompel<sup>®</sup> [14] database. CPModule in turn uses a very strict TFBSs prediction threshold and applies some additional filtering steps afterwards and the results of CrmMiner are restricted to enriched TFBSs. Although the breast cancer gene set is much more specific in contrast to the genome-wide gene set, the number of predicted pairs appears not to be smaller throughout all methods.

I performed a pairwise overlapping analysis of the identified pairs between all approaches (see Table 5.2). As expected, the largest overlap was found between the different distance

constraints of my approach. This is self-explaining since all pairs of maximal distance 20bp are included in the pairs of maximal distance 50bp and 100bp. All other pairwise combinations show a rather small overlap. This in turn results from the different information the methods focus on indicating that the approaches can perfectly complement each other.

**Table 5.2.: Pairwise comparison of the different approaches.** Number of predicted overlapping pairs of my approach (PC-TraFF) with three different maximal distances (20bp, 50bp and 100 bp), MatrixCatch (MC), CrmMiner (CrmM) and CPModule (CPM) for the breast cancer as well as for the genome wide dataset.

	Genome-wide analysis	Breast cancer analysis
$ PC-TraFF_{20} \cap PC-TraFF_{50} $	43	54
$ PC-TraFF_{20} \cap PC-TraFF_{100} $	41	43
$ PC-TraFF_{20} \cap MC $	3	1
$ PC-TraFF_{20} \cap CPM $	6	0
$ PC-TraFF_{20} \cap CrmM $	0	0
$ PC-TraFF_{50} \cap PC-TraFF_{100} $	82	80
$ PC-TraFF_{50} \cap MC $	4	1
$ PC-TraFF_{50} \cap CPM $	8	1
$ PC-TraFF_{50} \cap CrmM $	2	0
$ PC-TraFF_{100} \cap MC $	4	1
$ PC-TraFF_{100} \cap CPM $	9	0
$ PC-TraFF_{100} \cap CrmM $	2	0
$ MC \cap CPM $	1	0
$ MC \cap CrmM $	0	1
$ CPM \cap CrmM $	3	1

In order to statistically evaluate the performance of my approach and the three existing methods, I constructed all possible pairwise combinations of the PWMs in the library. I defined a pair to be part of the positive control set, if the interaction of the underlying TFs is experimentally validated and presented in TRANSCompel<sup>®</sup>, STRING [51] and BioGRID [52] interaction databases. All of the remaining pairs are defined to form the negative controls. Using both, negative and positive control sets, I calculated the sensitivity, specificity and Matthews correlation coefficient (MCC) for each method (see Table 5.3). It points out that all methods have a high specificity and a low sensitivity, indicating that the general performance of the methods is similar. The low sensitivity of all metrics can be explained by the negative control set that is overestimated in its size due to its definition, since some



of its TFs may in fact form functional yet undiscovered TF pairs. In general, my method has the highest sensitivity compared to the existing methods. Further, using larger maximal distance values seems to improve the sensitivity of my approach. The sensitivity of MatrixCatch, CPModule and CrmMiner appears to be similar with values between 0.5% and 0.6%. Regarding the specificity, CPModule reaches 100%, all other methods perform similar with specificity values between 99.3% and 99.9%. The Matthews correlation coefficient (MCC) is a measure for the general method performance and varies between -1 and +1. MCC-value is equal to 1 if the underlying model correctly classifies all test data, it is equal to 0 for a random prediction and in turn, equal to -1 if it classifies contradictory to the input data. All methods have a low MCC that can (similar to the sensitivity) be explained by the definition of the negative control set. While there should not be put too much weight on the absolute numbers of the MCC (as for sensitivity), comparison of the different methods clearly shows that my approach exhibits the highest MCC-values.

**Table 5.3.: Performance comparison of the different approaches.** Comparison of my approach with different maximal distance constraints ( $PC-TraFF_{20}$ ,  $PC-TraFF_{50}$ ,  $PC-TraFF_{100}$ ), MatrixCatch (MC), CPModule (CPM) and CrmMiner (CrmM).

	Sensitivity	Specificity	MCC
$PC-TraFF_{20}$	2.3%	99.5%	0.088
$PC-TraFF_{50}$	3.1%	99.3%	0.1
$PC-TraFF_{100}$	3.2 %	99.3%	0.102
MC	0.5 %	99.9 %	0.053
CPM	0.5%	100%	0.06
CrmM	0.6 %	99.6%	0.025

As mentioned before the pairwise comparison study reveals that the methods perfectly complement each other in their predictions. Therefore, I further analyzed the union of the predicted pairs to make a recommendation which method combination performs best. As shown in Table 5.4, the best sensitivity was reached using my approach with a maximal distance of 100bp in combination with the three other methods. The highest MCC-value (MCC=0.12) is reached by the union of the predicted pairs of  $PC-TraFF_{100}$ , MatrixCatch and CrmMiner. Following these results, I recommend the joint usage of all approaches in order to receive a large variation of cooperating TFs in the sequences under study.

**Genome-wide analysis** For the biological evaluation of my approach, I applied it to the genome-wide human RefSeq gene set that is linked to 23015 unique promoter sequences. I decided to use a maximal distance threshold of 20bps since the number of pairs is of

**Table 5.4.: Combination of the different approaches.** The general performance of the different methods can be improved by complementing the results of the different methods: my approach with maximal distances 20bp, 50bp and 100pb (*PC-TraFF<sub>20</sub>*, *PC-TraFF<sub>50</sub>*, *PC-TraFF<sub>100</sub>*), MatrixCatch (MC), CPModuel (CPM) and CrmMiner (CrmM).

	Sensitivity	Specificity	MCC
<i>PC-TraFF<sub>20</sub> ∪ MC</i>	2.8%	99.5%	0.101
<i>PC-TraFF<sub>50</sub> ∪ MC</i>	3.6%	99.3%	0.112
<i>PC-TraFF<sub>100</sub> ∪ MC</i>	3.8%	99.3%	0.114
<i>PC-TraFF<sub>20</sub> ∪ CPM</i>	2.6%	99.5%	0.099
<i>PC-TraFF<sub>50</sub> ∪ CPM</i>	3.4%	99.3%	0.107
<i>PC-TraFF<sub>100</sub> ∪ CPM</i>	3.5%	99.3%	0.109
<i>PC-TraFF<sub>20</sub> ∪ CrmM</i>	3.0%	99.2%	0.087
<i>PC-TraFF<sub>50</sub> ∪ CrmM</i>	3.8%	99%	0.10
<i>PC-TraFF<sub>100</sub> ∪ CrmM</i>	3.9%	99%	0.102
<i>MC ∪ CPM</i>	1.0%	99.9%	0.079
<i>MC ∪ CrmM</i>	1.2%	99.6%	0.05
<i>CPM ∪ CrmM</i>	1.2%	99.6%	0.051
<i>PC-TraFF<sub>20</sub> ∪ MC ∪ CPM</i>	3.1%	99.5%	0.11
<i>PC-TraFF<sub>50</sub> ∪ MC ∪ CPM</i>	3.8%	99.3%	0.118
<i>PC-TraFF<sub>100</sub> ∪ MC ∪ CPM</i>	4%	99.3%	0.12
<i>PC-TraFF<sub>20</sub> ∪ MC ∪ CPM ∪ CrmM</i>	3.8%	99.2%	0.10
<i>PC-TraFF<sub>50</sub> ∪ MC ∪ CPM ∪ CrmM</i>	4.5%	99%	0.116
<i>PC-TraFF<sub>100</sub> ∪ MC ∪ CPM ∪ CrmM</i>	4.7%	99%	0.119
<i>MC ∪ CPM ∪ CrmM</i>	1.7%	99.6%	0.07

manageable size for a manual evaluation and the performance difference to the other tested maximal distances is negligible. Using a maximal distance threshold of 20bp I predicted 54 significant TFBS pairs. Considering the underlying TF pairs, seven pairs refer to homotypic interactions while the remaining 47 pairs form heterotypic interactions. In total, 44 pairs are experimentally validated and described in TRANSCompe<sup>®</sup>, BioGRID or STRING database. The remaining ten pairs are not described in literature yet and are either false positive predictions or present new targets for experimental validation. The top ten TFBS pairs according to *z-score* ranking are given in Table 5.5 in combination with literature reference if available.

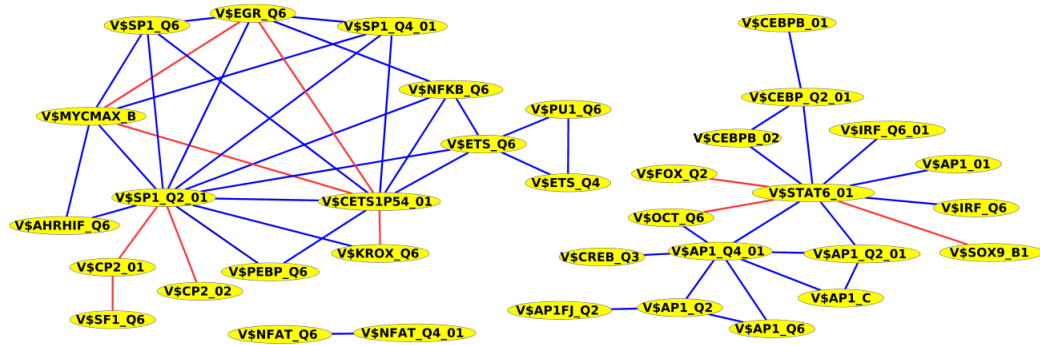
**Table 5.5.: Significant TFBS pairs found by the method in genome-wide promoter analysis of human RefSeq genes.** The table shows the top 10 significant TFBS pairs, which are sorted in descending order based on their *z-scores* in combination with their literature evidence in BioGRID, STRING and TRANSCompel<sup>®</sup>

Significant Pair	<i>z-score</i>	Reference
V\$PU1_Q6 - V\$ETS_Q6	9.84	TRANSC, BioGRID, STRING
V\$CETS1P54_01 - V\$ETS_Q6	5.76	TRANSC, BioGRID, STRING
V\$ETS_Q4 - V\$ETS_Q6	5.49	TRANSC, BioGRID, STRING
V\$EGR_Q6 - V\$SP1_Q2_01	5.09	BioGRID, STRING
V\$CETS1P54_01 - V\$SP1_Q2_01	4.94	TRANSC, STRING
V\$AP1_Q2_01 - V\$AP1_Q4_01	4.69	TRANSC, BioGRID
V\$STAT6_01 - V\$OCT_Q6	4.66	-
V\$CEBPB_02 - V\$STAT6_01	4.58	TRANSC, STRING
V\$MYCMAX_B - V\$SP1_Q2_01	4.36	BioGRID, STRING
V\$AP1FJ_Q2 - V\$AP1_Q2	4.09	TRANSC, BioGRID, STRING

\*TRANSC:TRANSCompel<sup>®</sup>

For a visual analysis of the results, I constructed cooperation networks based on the predicted TFBS pairs, where the nodes correspond to TFBSs and edges to predicted cooperations between them. The cooperation network of the whole genome analysis is shown in Figure 5.1 and consists of 35 nodes and 54 edges. The network is separated into two large unconnected subgraphs of similar size regarding the number of nodes and the binding sites V\$NFAT\_Q6 and V\$NFAT\_Q4\_01 that form an isolated TFBS pair. It can be seen that the majority of the TFBS motifs of the left cluster consists of GC-rich sequences while those of the right cluster tend to contain more AT-rich patterns (see Table 5.6), which is in accordance with the findings that Hu et al. made in their study in 2007 [8]. For the entire network four hub nodes can be identified: V\$SP1\_Q2\_01, V\$STAT6\_01, V\$CETS1P54\_01 and V\$AP1\_Q4\_01, that are presented in Table 5.7 in combination with their top three interaction partners according to *z-core* ranking and a literature evidence if available.

The first hub is V\$SP1\_Q2\_01 that is part of twelve significant pairs. Of these pairs, ten can be confirmed by experimental studies and two of them (V\$SP1\_Q2\_01 - V\$CP2\_01 and V\$SP1\_Q2\_01 - V\$CP2\_01) are new pairs. This hub is placed in the left subgraph of the network and represents a cytosine-rich motif. The binding site V\$SP1\_Q2\_01 is bound by factor SP1 which is a member of the three zinc finger Krüppel-related transcription factor family [25] that is known to bind to GC-rich promoter regions [53] and promoters that lack



**Figure 5.1:** Cooperation network of PC-TraFF significant TFBS pairs of whole genome analysis. The nodes of the network refer to transcription factor binding sites and the edges between them to a predicted pairing. Blue edges indicate an experimentally validated interaction while the red ones are newly predicted interactions. (Figure from [43])

a TATA box [54]. On the functional side of view, the transcriptional activity of SP1 is important for the activation of a multitude of housekeeping genes [54] and regulates genes involved in cell proliferation, apoptosis, differentiation and neoplastic transformation [53]. For the latter case, SP1 is responsible for the transcriptional initiation step by recruiting the transcriptional machinery [54]. In general, SP1 can act as a transcriptional activator or repressor dependent on the target promoter and the co-factors or TFs it interacts with [53].

**Table 5.6:** Exemplary comparison between the TFBSs contained in the left and the right cluster of the cooperation network of the whole genome analysis. While the TFBSs of the left cluster are more GC-rich, those of the right cluster are AT-rich.

Left cluster	Right cluster
V\$SP1_Q2_01	V\$FOX_Q2
V\$CETS1P54_01	V\$IRF_Q6
V\$MYCMAX_B	V\$AP1_C
V\$EGR_Q6	V\$CEBPB_01
V\$VCP2_02	V\$SOX9_B1

The second hub is V\$STAT6\_01 that is paired to ten other TFBSs. Considering the underlying TFs, three pairs are novel while the remaining pairs are experimentally validated and described in literature. V\$STAT6\_01 is bound by factor STAT6, a member of the *signal transducer and activator of transcription* (STAT) family. STAT6 is known to act in

response on cytokines IL-4 and IL-13 and thus, it is involved in the immune system [55]. Besides its function in T-cells and B-cells, STAT6 is linked to cellular processes in the mammary gland, lung and skin [55].

The next hub is V\$CETS1P54\_01 which is paired to nine other TFBSs, three of which appear to be novel pairs and six have already been described in literature. V\$CETS1P54\_01 represents the GC-rich binding site of factor ETS1 and is located in the left GC-rich subgraph of the cooperation network. ETS1 is a member of the ETS transcription factor family that is known to be crucial for the regulation of normal cell proliferation and differentiation [56].

The last hub is V\$AP1\_Q4\_01 and is located in the right subgraph of the network, even though the amount of A,T,G and C appears to be balanced in that motif. V\$AP1\_Q4\_01 is involved in seven pairings that are all experimentally confirmed. The binding site is bound by factor AP1, a dimerized protein of members of the basic leucine zipper factor family [25]. This dimerization is either a homotypic or heterotypic linking between members of JUN-related factors, FOS-related factors and/or activating transcription factors (ATFs) [57, 58]. Thus, a lot of different dimerization combinations are possible that are all referred to AP1 molecule and in dependence of the protein combinations the functions of AP1 differ. In general, AP1 is involved in cellular processes such as proliferation, differentiation, apoptosis and transformation [57].

The hub nodes V\$SP1\_Q2\_01 and V\$CETS1P54\_01 form a significant TFBS pair. The interaction between the corresponding TFs SP1 and ETS1 is known for some promoters lacking a TATA box where SP1 can replace the functionality of the TATA box, since the binding site for SP1 is of low affinity but is strengthened by the adjacent binding of ETS1 [59].

Further, the hubs V\$STAT6\_01 and V\$AP1\_Q4\_01 are predicted to form a significant pair. The underlying TF interaction of STAT6 and JUN plays an important role in the activation process of the IL-24 promoter. IL-24 in turn is crucial for B cell differentiation and anticancer processes in a variety of diverse cancer cells [60].

**Breast cancer gene set analysis** In order to test the methodology to a more specific gene set, I applied it to 218 promoter sequences determined for the breast cancer associated Ref-Seq gene set. Following the proceeding of the whole genome analysis, I chose a maximal distance threshold of 20bp between the TFBSs.

For these breast cancer related promoter regions, I identified 64 significant TFBS pairs of which 5 pairs can be linked to homotypic TF interactions and the remaining ones to heterotypic TF interactions. Comparing the pairs with known TF interactions, 44 pairs are published in protein interaction databases like STRING, BioGRID and TRANSCompel<sup>®</sup>,

**Table 5.7.: The hubs and their top three collaboration partners** in the predicted collaboration network of significant TFBS pairs for human RefSeq genes and their literature evidence in BioGRID, STRING and TRANSCompel<sup>®</sup>.

Hub	Top three collaborating partners	z-score	Reference
V\$SP1_Q2_01	V\$EGR_Q6	5.09	BioGRID, STRING
	V\$CETS1P54_01	4.94	TRANSC, STRING
	V\$MYCMAX_B	4.36	BioGRID, STRING
V\$STAT6_01	V\$OCT_Q6	4.66	-
	V\$CEBPB_02	4.58	TRANSC, STRING
	V\$CEBP_Q2_01	3.74	TRANSC, BioGRID, STRING
V\$CETS1P54_01	V\$ETS_Q6	5.76	TRANSC, BioGRID, STRING
	V\$SP1_Q2_01	4.94	TRANSC, STRING
	V\$NFKB_Q6	3.96	TRANSC, STRING
V\$AP1_Q4_01	V\$AP1_Q2_01	4.69	TRANSC, BioGRID, STRING
	V\$STAT6_01	3.35	TRANSC, BioGRID, STRING
	V\$AP1_Q6	3.35	TRANSC, BioGRID, STRING

\*TRANSC:TRANSCompel<sup>®</sup>

whereas 20 pairs are novel predictions and provide new targets for experimental approaches.

The breast cancer cooperation network (see Figure 5.2) consists of 40 nodes, representing TFBSs and 64 edges that refer to predicted cooperations between the related TFBSs. The network is composed of two large unconnected subgraphs and two separate node pairs, where the subgraphs consist of 10 and 16 nodes, respectively. Having a closer look at the TFBS representing logoplots of the two major clusters, it can be seen that the upper cluster in Figure 5.2 contains more GC-rich motifs while the lower cluster has more AT-rich clusters. However, this trend is not as distinctive as for the whole genome analysis.

For the breast cancer cooperation network, three hubs can be identified: V\$NFKB\_Q6, V\$CETS1P54\_01 and V\$MYCMAX\_B.

The hub node V\$CETS1P54\_01 is involved in thirteen TFBS pairings where eight of them are confirmed by literature and the remaining five pairs appear to be new targets for validation experiments. V\$CETS1P54\_01 is bound by ETS1 that regulates genes involved in the regulation of tumor progression and metastasis in breast cancer cells [61].

**Table 5.8.: The hubs and their top three collaboration partners** in the predicted collaboration network of breast cancer-associated significant TFBS pairs for human RefSeq genes and their literature evidence in BioGRID, STRING and TRANSCompel<sup>®</sup>.

Hub	Top three collaborating partners	z-score	Reference
V\$NFKB_Q6	V\$CETS1P54_01	5.42	TRANSC, STRING
	V\$ETS_Q6	4.80	BioGRID, TRANSC, STRING
	V\$SP1_Q4_01	3.43	BioGRID, TRANSC, STRING
V\$CETS1P54_01	V\$ETS_Q6	8.01	BioGRID, TRANSC, STRING
	V\$NFKB_Q6	5.42	TRANSC, STRING
	V\$MYC_MAX_B	5.21	-
V\$MYC_MAX_B	V\$CETS1P54_01	5.16	-
	V\$E2F_Q3_01	5.21	TRANSC
	V\$AHRHIF_Q6	4.39	BioGRID, STRING

\*TRANSC:TRANSCompel<sup>®</sup>

V\$NFKB\_Q6 is paired with ten other TFBSs, all of which are related to experimentally validated TF interactions. The binding site is bound by the NF- $\kappa$ B transcription factor family. These factors are known to be involved in cell proliferation, survival, immunity, inflammation regulation and angiogenesis [62]. Moreover, it has been detected to be involved in breast cancer [63]. In this study, I found NF- $\kappa$ B to cooperate with the factors ETS1, ELF1, SP1 and E2F1, each of which is linked to breast cancer. In detail, ETS1 is under suspicion to be a breast cancer oncogene by regulating tumor progression and metastasis [61]. ELF1 is also a member of the ETS transcription factor family and is in general linked to the regulation of cellular growth and differentiation [64], however, the up-regulation of the ELF1 gene has been detected in prostate and breast cancer cells [64].

The third hub is V\$MYC\_MAX\_B that is paired to nine other TFBSs. Three pairings are novel targets for laboratory experiments while the others have already been described in literature. V\$MYC\_MAX\_B is bound by the heterodimer of factors MYC and MAX where the dimerization process is important for the activation of the MYC protein [65]. MYC is in general important for the regulation of cell growth, proliferation, metabolism, differentiation and apoptosis [66]. Further, it is known to be involved in breast cancer [66, 67].

A closer look at the cooperation network shows that two TFBSs (V\$E2F\_Q3\_01 and V\$E2F\_Q4\_01) are related to the E2F transcription factor family. This family is involved

in cell cycle regulation, apoptosis as well as DNA damage response [68]. E2F family members are known to be involved in breast cancer, more specifically, in the down-regulation of BRCA1 gene expression in response to hypoxia. This in turn is mediated by two E2F factors binding close to each other to the BRCA1 promoter [68].

Three TFBSs (V\$CEBPB\_02, V\$CEBP\_Q2 and V\$CEBP\_Q2\_01) in the network of significant pairs can be linked to C/EBP $\beta$ . Members of the C/EBP $\beta$  family are known to be involved in cellular functions tied to tumor progression like proliferation, survival and apoptosis and are further linked to malignant transformation of human breast [69]. According to my analysis, the factor is among others interacting with HMGA1 and c-Myb. The high mobility group A1 (HMGA1) factor is found to be enriched in embryonic tissues and differentiated tumors [70], whereas the factor c-Myb plays a critical role in the regulation of cell cycle and has been identified to be involved in carcinoma, breast cancer and colon cancer [71].

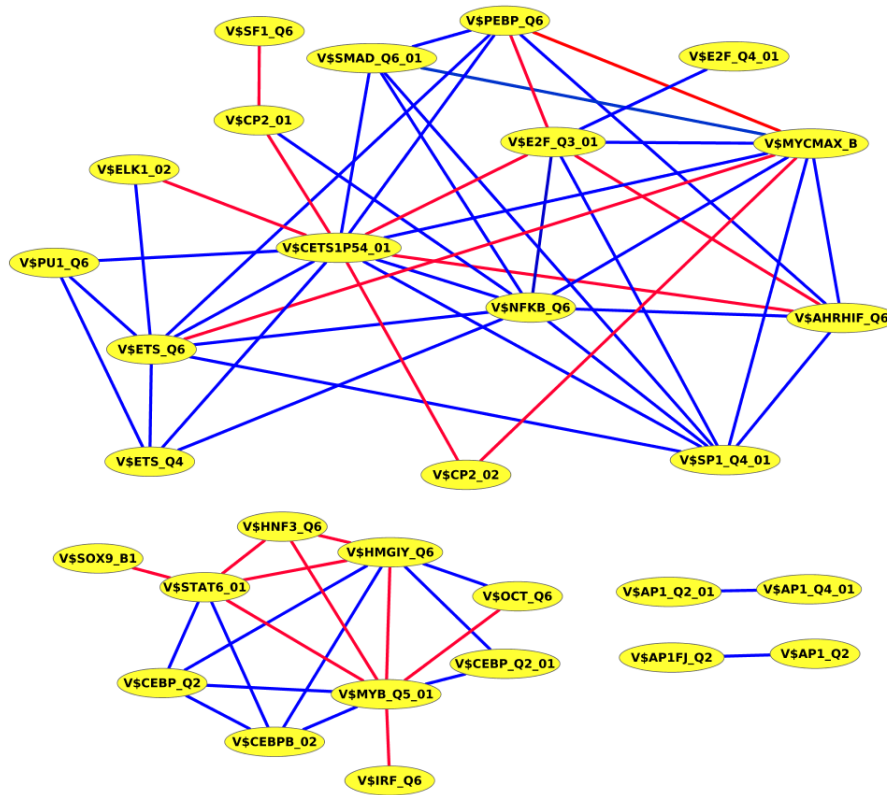
### 5.1.2. Sequence-set specific cooperating TFs

In this section I present the results of the application of the extended version (see Section 4.1.2) in a comparative manner to the original approach (see Section 4.1.1) in order to evaluate its performance in the separation of sequence-set specific intra-regional TF cooperations from the generally important ones.

**Datasets** In order to extensively evaluate the performance of the approach, I tested it in two different ways. First, I used the simulation dataset (see Section 5.1.1) where I artificially inserted the TFBS pair (V\$USF\_01-V\$IRF1\_01). Second, I analyzed gene sets obtained from Joshi et al. [49] of five breast cancer associated subtypes: Luminal A, Luminal B, Basal-like, Normal-like and ErbB2 over-expressing. For these gene sets I selected the associated promoter sequences in the range -500bp to +100bp relative to transcription start site (TSS). The results show that the numbers of genes and consequently the numbers of promoter sequences under analysis strongly differ between the individual subtypes (see Table 5.9), which enables us to further assess the performance of the approach regarding the size of different input sets.

**Analysis of simulation dataset** In order to evaluate the performance of the extended approach (see Section 4.1.2), I tested it on a simulation dataset of 200 sequences where I randomly inserted the binding site pair of V\$USF\_01-V\$IRF1\_01. The analysis of this dataset with the original approach results in 58 significant TFBS pairs where the inserted pair is on position 18 in the *z-score* ranking. Applying the extended approach in the simplest version, I subtracted the calculated background cooperation level from the initial  $\text{PMII}$ -values. This results in 55 specific TFBS pairs and three TFBS pairs were identified as common (generally important) ones, thereby the inserted pair is raised onto position 16 in *z-score* ranking.





**Figure 5.2.: Cooperation network of PC-TraFF significant TFBS pairs of breast cancer gene set analysis.** The nodes of the network refer to transcription factor binding site types and the edges between them to a predicted pairing. Blue edges indicate an experimentally validated interaction while the red ones are newly predicted interactions. (Figure from [43])

**Table 5.9.: Number of promoter sequences of breast cancer subtype-associated RefSeq genes and corresponding significant pairs found by my approach.**

Subtype	Number of genes	Number of promoter sequences
Luminal A	78	86
Luminal B	55	57
Basal-like	28	31
Normal-like	23	27
ErbB2 over-expressing	13	15

The low number of common pairs indicates that the quantification of a background level can be difficult for unspecific sequence-sets and consequently, the separation of sequence-set specific cooperations from common ones might not be possible. In order to overcome this problem to some extent, I further introduced the parameter  $\alpha$  in order to scale the subtracted background and at the same time to increase the specificity of the predicted TFBS pairs.

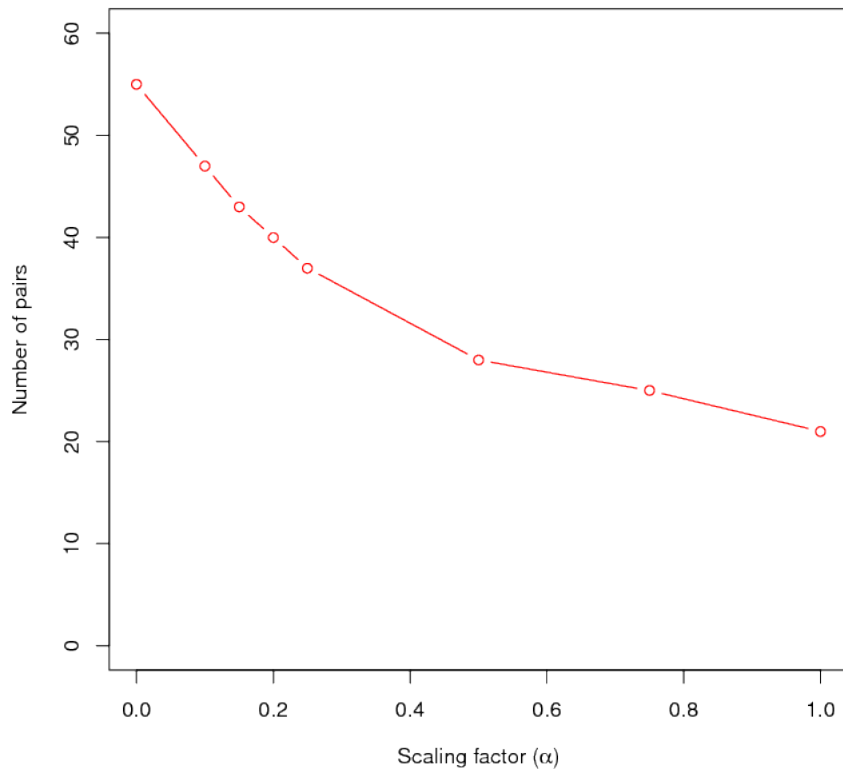
In the following, the pairs identified with the original approach are still referred to as significant pairs, whereas those pairs among them identified as sequence-set specific pairs by the extended version are referred to as specific pairs.

**Table 5.10.: Total number of specific TFBS pairs for the simulation dataset for different  $\alpha$ -values.** The rank of the inserted pair gives the position of the inserted pair according to  $z$ -score ranking.  $\alpha=-1$  indicates the significant pairs identified by the original method.

$\alpha$ -value	Rank of inserted pair	Total number of pairs
-1	18	58
0	16	55
0.1	15	47
0.15	14	43
0.2	12	40
0.25	11	37
0.5	6	28
0.75	6	25
1	5	21

The influence of the scaling factor  $\alpha$  on the number of predicted specific pairs is shown in Figure 5.3. Setting  $\alpha = -1$  refers to the predictions of the original approach, whereas setting  $\alpha = 0$  results in subtracting the estimated background level from the initially calculated  $\mathbb{P}\text{MII}$ -values and in turn  $\alpha = 1$  results in the subtraction of the doubled  $\mathbb{P}\text{MII}$  mean value. It can be seen that the number of specific pairs decreases with an increasing  $\alpha$ -value from 55 pairs ( $\alpha = 0$ ) to 21 pairs for  $\alpha = 1$ . A closer look at Figure 5.3 indicates that the influence of  $\alpha$  on the number of significant pairs is not linear, although  $\alpha$  has a linear influence on the subtracted background level. It has to be noted that the inserted pair was identified for all  $\alpha$ -values as sequence-set specific pair.

The position of the inserted pair in the  $z$ -score ranking is rising in accordance to an increasing  $\alpha$ -value. Starting on position 18 in the original analysis, it ends up on position five for  $\alpha = 1$ , indicating that the specificity of the predicted pairs is increased. However, the



**Figure 5.3.:** Number of specific TFBS pairs in dependence on different  $\alpha$ -values for the simulation dataset.

inserted pair is not on the first position in the  $z$ -score ranking as it might be expected. This can be explained by a closer look at the logoplots of the underlying TFBSs of the top four ranked TFBS pairs (V\$PU1\_Q6-V\$ETS\_Q6, V\$IRF1\_01-V\$TAXCREV\_01, V\$HNF4\_Q6-V\$GR\_Q6\_01 and V\$IRF1\_01-V\$ATF3\_Q6) setting  $\alpha = 1$  that either contain one of the inserted TFBSs or are very similar to them (see Figure 5.4). Thus, these TFBSs tend to match the inserted sequences as well, due to their sequence similarity and therefore I unintentionally incorporated these TFBS pairs in the sequences which in turn classifies them as true positive pairs. The TFBS pair V\$PU1\_Q6-V\$ETS\_Q6 is top ranking for all  $\alpha$ -values. Although both TFBSs only match the consensus sequence of V\$IRF1\_01 they are in general very short and unspecific which results in many sequence matches. This number of matches is increased by the insertion of V\$IRF1\_01 consensus sequence leading to a huge overestimation of the V\$PU1\_Q6-V\$ETS\_Q6 pair that can not be dropped down by background

subtraction.

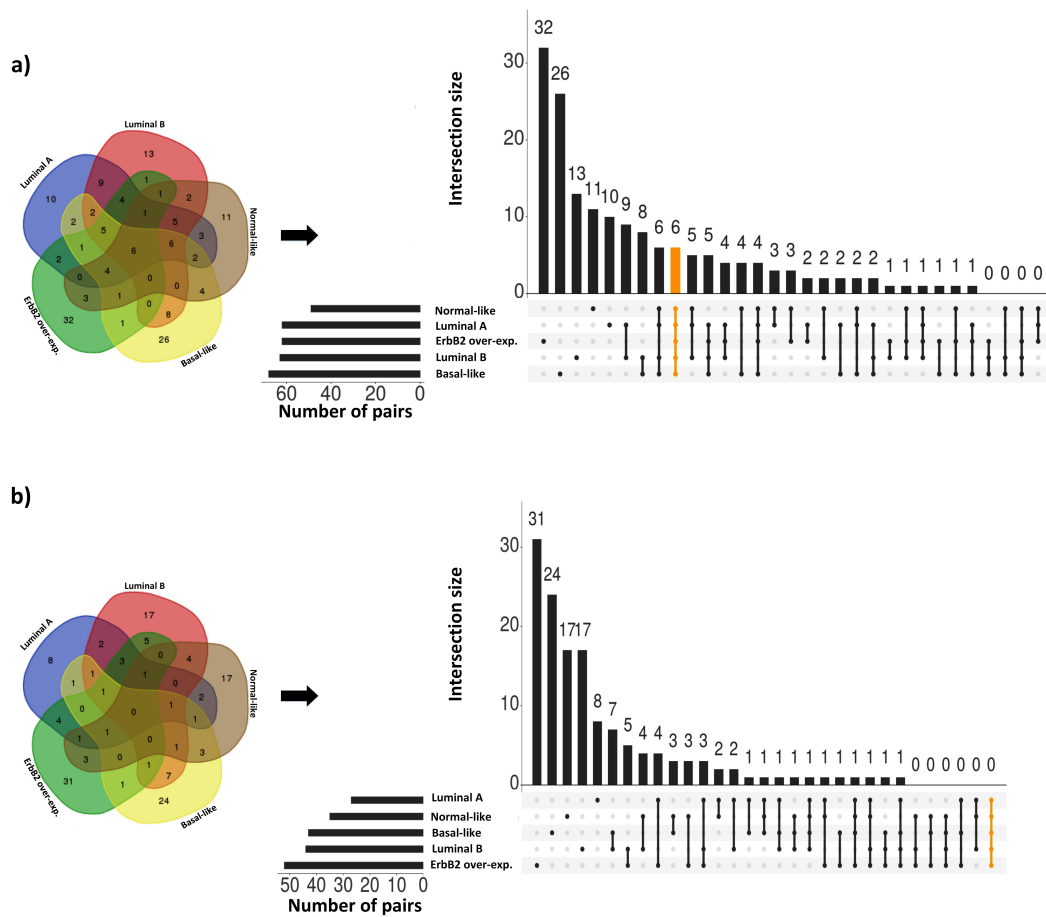


**Figure 5.4.: Logoplot alignment for the TFBSs involved in the four top ranking pairs** (V\$PUI\_Q6-V\$SETS\_Q6, V\$IRF1\_01-V\$TAXCREV\_01, V\$HNF4\_Q6-V\$GR\_Q6\_01 and V\$IRF1\_01-V\$ATF3\_Q6) setting  $\alpha = 1$  of the simulation dataset where I artificially inserted the TFBS pair V\$USF\_01-V\$IRF1\_01.

**Breast cancer subtypes** In order to proof the gained specificity of the extended approach in comparison to the original method, I performed a comparison study of the original approach and its extended version for five breast cancer subtypes: Luminal A, Luminal B, Normal-like, Basal-like and ErbB2 over-expressing.

Applying the original approach to the promoter sequences of the five breast cancer gene sets I detected 62 significant TFBS pairs for Luminal A, 63 pairs for Luminal B, 68 pairs for Basal-like, 49 pairs for Normal-like and 62 significant TFBS pairs for ErbB2 over-expressing. Comparing the significant pairs of the different subtypes with each other shows that there is a large overlap among these pairs and six pairs are identified as significant in all subtypes (see Figure 5.5a)) although the underlying gene sets are unique for each subtype. The reason for this large overlap of significant pairs might stem from the common regulatory programs present in all cells in general or in cells sharing a common origin. However, the differentiation of these generally important pairs from the sequence set specific ones is not possible for the original approach.

In order to determine the sequence set-specific TFBS pairs for each breast cancer subtype, I applied the extended approach to the promoter sequences of the underlying gene sets. I varied  $\alpha$  in order to estimate its influence on the results (see Figure 5.6). By subtracting the mean  $\mathbb{P}_{\text{MII}}$ -value (setting  $\alpha=0$ ) the extension approach categorizes on average 90% of the significant TFBS pairs to be specific for the respective sets. Increasing  $\alpha$  leads to a reduction of the specific pairs. However, the strength of this reduction differs between the individual subtypes (see Figure 5.6). For Luminal A subtype, the number of specific pairs drops dramatically by increasing  $\alpha$  and finally, by setting  $\alpha = 1$  about 1% of the significant pairs is predicted to be sequence set specific for Luminal A breast cancer subtype. In contrast, the number of ErbB1 sequence-set specific pairs is only slightly decreasing in dependence of  $\alpha$  and 47% of the significant pairs are identified as sequence-set specific for  $\alpha = 1$ .



In Figure 5.5b I exemplarily present the numbers of unique and overlapping specific TFBS pairs of the breast cancer subtypes using  $\alpha = 0.2$ . Regarding these results, in the original approach, Luminal A shows the smallest number of unique pairs. Interestingly, the number of unique pairs of Luminal B subtype as well as of Normal-like subtype is raised in comparison to the original analysis. For Normal-like subtype, there are eleven significant unique pairs (Figure 5.5a) and 17 specific unique pairs (Figure 5.5b). Thus, there are

**Table 5.11.: Six significant TFBS pairs determined as significant by the original approach for all breast cancer subtypes.**

Significant pairs	Reference
V\$MYCMAX_B - V\$E2F_Q3_01	TRANSCompel <sup>®</sup>
V\$CETS1P54_01 - V\$PEBP_Q6	TRANSCompel <sup>®</sup> , BioGRID, STRING
V\$CETS1P54_01 - V\$NFKB_Q6	TRANSCompel <sup>®</sup> , STRING
V\$CEBP_Q2 - V\$STAT6_01	TRANSCompel <sup>®</sup> , BioGRID, STRING
V\$AP1_Q2_01 - V\$AP1_Q4_01	TRANSCompel <sup>®</sup> , BioGRID, STRING
V\$CEBPB_02 - V\$STAT6_01	TRANSCompel <sup>®</sup> , STRING

six pairs solely sequence-set specific for Normal-like subtype that have also been predicted as significant by the original approach in other subtypes (see Table 5.12). For example, the pairs V\$CEBPB\_02-V\$HMGY\_Q6 and V\$ELK1\_02-V\$CETS1P54\_01 are significant for four subtypes, respectively, but specific only for Normal-like. In addition, the TFBS pair V\$EGR\_Q6-V\$AHRHIF\_Q6 is significant for Basal-like and Normal-like but only specific for Normal-like subtype. For Luminal B subtype, 13 pairs are identified as significant unique according to the original methodology and 17 pairs are unique specific for this subtype. Of this, six pairs have been predicted to be significant for several subtypes (see Table 5.12). In addition, three of the 13 pairs that are significant unique for Luminal B (V\$MYB\_Q5\_01-V\$MAF\_Q6\_01, V\$NFKB\_Q6-V\$CP2\_02 and V\$HMGY\_Q6-V\$MAF\_Q6\_01) have not been identified as specific for this subtype.

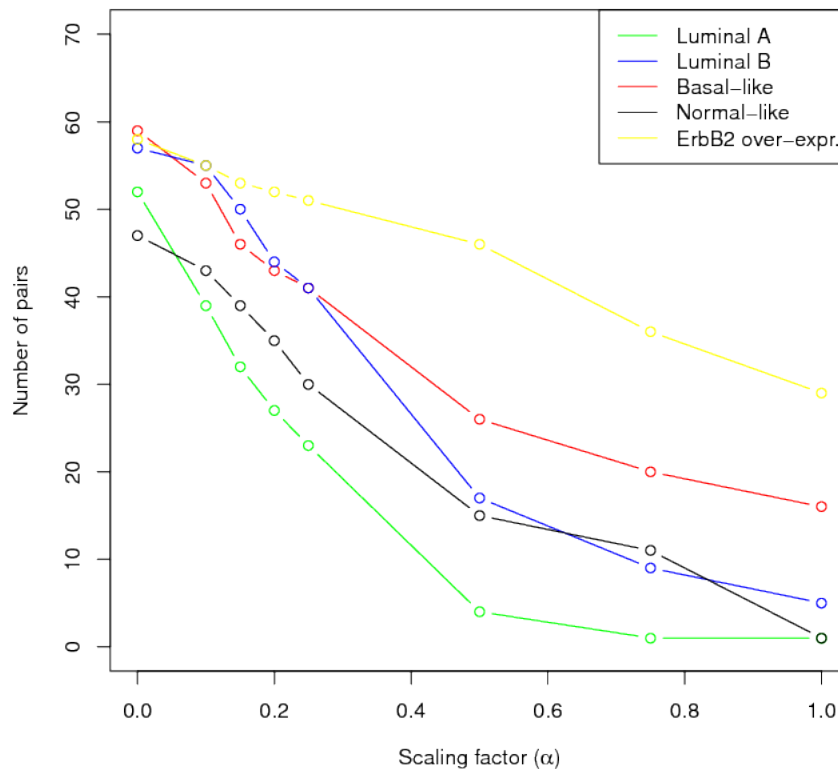
Regarding the overlap of sequence-set specific pairs using  $\alpha = 0.2$  the number of overlapping pairs has been decreased dramatically in comparison to the original approach. In particular, six pairs have been identified as significant for all BRC subtypes in the original analysis (see Figure 5.5a), but there is no pair predicted as specific in all subtypes (see Figure 5.5b). These six overlapping pairs of the original approach are listed in Table 5.11 and are predicted to be specific for some of the breast cancer subtypes but not for all of them. For example the pair V\$CETS1P54\_01 - V\$NFKB\_Q6 is predicted to be specific for Luminal A, Normal-like and Basal-like subtype whereas the pairs V\$MYCMAX\_B - V\$E2F\_Q3\_01 and V\$STAT6\_01 - V\$HMGY\_Q6 are uniquely identified as specific for Basal-like and Normal-like, respectively.

As for the original method, I built cooperation networks based on the significant pairs where nodes refer to TFBSs and edges to predicted pairings between them. Figure 5.7 shows the cooperation network of Luminal A subtype that consists of 33 nodes and 62 edges and is based on the significant TFBS pairs of the original approach. By only considering the specific pairs (using  $\alpha = 0.2$ ) seven nodes and 35 edges are eliminated from the original

**Table 5.12.: Pairs that were identified as significant** by the original method ( $\alpha = -1$ ) for different BRC-subtypes but are specific solely for a certain subtype using  $\alpha = 0.2$  for the background correction.

Specific for subtype	TFBS pairs	Significant in subtypes
Normal-like	V\$CEBPB_02 - V\$HMGY_Q6	<i>Basal-l., Luminal A, Luminal B, Normal-l.</i>
	V\$SELK1_02 - V\$CETS1P54_01	<i>Basal-l., Luminal A, Luminal B, Normal-l.</i>
	V\$CEBPB_02 - V\$CEBP_Q2	<i>ErbB2 over-expr., Luminal B, Normal-l.</i>
	V\$NFKB_Q6 - V\$SP1_Q4_01	<i>Luminal A, Normal.</i>
	V\$EGR_Q6 - V\$AHRHIF_Q6	<i>Basal-l., Normal-l.</i>
	V\$GR_Q6_01 - V\$PR_Q2	<i>ErbB2 over-expr., Normal.</i>
Luminal B	V\$CETS1P54_01 - V\$AHRHIF_Q6	<i>Luminal A, Luminal B, Normal-l.</i>
	V\$E2F_Q3_01 - V\$PEBP_Q6	<i>Luminal A, Luminal B</i>
	V\$MYC_MAX_B - V\$AHRHIF_Q6	<i>Basal-l., Luminal A, Luminal B</i>
	V\$NFKB_Q6 - V\$E2F_Q3_01	<i>Luminal A, Luminal B</i>
	V\$NFKB_Q6 - V\$AHRHIF_Q6	<i>Luminal A, Luminal B</i>
	V\$CETS1P54_01 - V\$CP2_02	<i>Luminal A, Luminal B</i>
V\$CETS1P54_01 - V\$MYC_MAX_B	<i>Basal-l., Luminal A, Luminal B, Normal-l.</i>	

network resulting in a network of 26 nodes and 27 edges. It is remarkable that the hubs of the original network of significant pairs developed totally different in comparison to the specific network by either maintaining a hub node, lost the property of being a hub, or totally vanished from the network. For example, the binding sites V\$CETS1P54\_01, V\$MYB\_Q5\_01 and V\$HMGY\_Q6 are still highly connected nodes in the specific cooperation network, although they lost some cooperation partners (neighbouring nodes). The nodes V\$NFKB\_Q6 and V\$AHRHIF\_Q6 are hub nodes in the significant pair network but become low connected nodes in the specific network. In turn, hub node V\$SP1\_Q4\_01 is totally missing in the specific cooperation network. In contrast to this, V\$SMAD\_Q6\_01 lost one of its neighbours and appears to be one of the highly connected nodes of the new network.



**Figure 5.6.: Number of sequence-set specific TFBS pairs for the five breast cancer subtypes in dependence of scaling factor  $\alpha$ .** (Figure from [44])

In Figure 5.8, I show the dynamic change of the cooperation network for Basal-like subtype in dependence of different  $\alpha$ -values. The network of significant pairs, resulting from the original approach, consists of 36 nodes and 68 edges. Transforming the network in a way that it is only build up by the specific pairs (setting  $\alpha = 0.2$ ) results in the elimination of 43 edges, while the number of nodes remains the same. A further increase in  $\alpha$  leads to a clear reduction of network size. In example, the network of specific pairs using  $\alpha = 0.5$  consists of only 25 nodes and 26 edges.

Comparing the network of Luminal A and Basal-like subtypes reveals that  $\alpha$  has a stronger influence on the Luminal A subtype network than on that of Basal-like subtype. A reason for this observation might stem from a more specific transcriptional regulation in Basal-like cells in comparison to Luminal A leading to a higher background cooperativity of TFs in Luminal A related promoter regions.







## 5.2. Identification of inter-regional associated TFs using multivariate mutual information

In this section I present the results for the identification of associated transcription factors between enhancer and their related promoter regions based on multivariate mutual information measures. To achieve this, I first constructed two simulation sets in order to evaluate and compare the performance of the different multivariate mutual information measures. Afterwards, I compared my approach with an existing method and analyzed the sequences of promoter-enhancer interactions (PEIs) of six different human cell lines for their underlying associated transcription factor pairs.

### 5.2.1. Example dataset

In order to illustrate my basic idea to the reader and to get a first impression about the general performance of the different multivariate mutual information metrics, I constructed two small TFBS count matrices, one for enhancer and one for promoter sequences (see Figure 5.9). One row in each matrix corresponds to an enhancer/promoter sequence, respectively, where both sequences correspond to a certain PEI. Columns represent the TFBS names and an entry in the matrix refers to the number of predicted PWM matches. The first four PEIs ( $E1/P1, \dots, E4/P4$ ) are defined as real/input PEIs, while the others are treated as background ( $E1_{sh}/P1_{sh}, \dots, E4_{sh}/P4_{sh}$ ). The type of the pairing  $L \in \{I, B\}$  is given in the vector  $\mathbb{V}^{lab}$ . Three TFBS types are predicted in the enhancer regions ( $T_E1, T_E2$  and  $T_E3$ ) and three in the promoters ( $T_P1, T_P2$  and  $T_P3$ ). Having three predictable binding site motifs in each, enhancer and promoter sequence, there are in total nine pairwise enhancer-promoter TFBS combinations.

A closer look at the count value distributions of the individual TFBSs provides hints about their general binding behaviour in the sequence set under study and provides a first insight about the pairwise association between two TFBSs of enhancer and promoter sequences. In this synthetic example, the binding behaviour of  $T_E1$  and  $T_P1$  appears to be associated in the real PEIs as well as in the background set. This pair is perfectly associated in my point of view, but it has to be pointed out that the modeled association in the background sequences is not likely to generate with my background set. The motif pair of  $T_E2$  and  $T_P2$  is associated in the real PEIs, but not in the background set and therefore, it appears to be the second best associated TFBS pair in this example. In contrast,  $T_E3$  and  $T_P3$  show an associated behaviour that is not related to the label (input or background) of PEI and refers to be the non-associated pair.

By assuming that the count matrices already contain the interval identifiers assigned in Phase 5, I applied the information theoretic measures to this example. Afterwards, I normalized the outcomes of the different quantities using the logarithm of the maximal alpha-

$M^{enh}$				$M^{prom}$				$\mathbb{V}^{lab}$
	$T_E1$	$T_E2$	$T_E3$		$T_P1$	$T_P2$	$T_P3$	$L$
E1	1	10	1	P1	2	1	1	I
E2	1	10	2	P2	2	1	2	I
E3	1	10	3	P3	2	1	3	I
E4	1	10	4	P4	2	1	4	I
$E1_{sh}$	4	9	1	$P1_{sh}$	9	2	1	B
$E2_{sh}$	4	1	2	$P2_{sh}$	9	3	2	B
$E3_{sh}$	4	0	3	$P3_{sh}$	9	4	3	B
$E4_{sh}$	4	2	4	$P4_{sh}$	9	5	4	B

**Figure 5.9.: Example dataset:** Synthetic generated TFBS-count matrices  $M^{enh}$  and  $M^{prom}$  and label vector  $\mathbb{V}^{lab}$ . The rows of  $M$  correspond to PEIs and the columns to TFBS names and an entry in the matrix is the frequency of predicted TFBSs in the respective sequence.  $\mathbb{V}^{lab}$  indicates the label of the interaction type (I refers to real/input PEIs, B indicates the background).

bet size in order to reduce the influence of alphabet size and to enable a proper comparison between the different quantities (see Table 5.13).

Using  $\mathbb{M}\mathbb{M}\mathbb{I}$ , the best associated pair  $T_P1-T_E1$  has the highest value with  $\mathbb{M}\mathbb{M}\mathbb{I}(T_E1; T_E2; L) = 1$ . The non-associated pair results in  $\mathbb{M}\mathbb{M}\mathbb{I}(T_E3; T_P3; L) = 0$ , indicating that the three variables  $T_E3$ ,  $T_P3$  and  $L$  do not contain any information about another. The second best associated pair gets a value of  $\mathbb{M}\mathbb{M}\mathbb{I}(T_E2; T_P2; L) = 0.43$  and thus, it is in the intermediate position.

Using the  $\mathbb{J}\mathbb{M}\mathbb{I}$ , I successfully identified the best associated pair as top ranking. The non-associated pair gets a value of 0, indicating that the joint distribution of  $T_P3 - T_E3$  and the label  $L$  do not share any commonality. The second best associated pair results in  $\mathbb{J}\mathbb{M}\mathbb{I}(T_E2, T_P2; L) = 0.43$  and is therefore on intermediate position of the three considered pairs. Having a look at the other potential pairings between the TFBSs of enhancer and promoter, respectively, it is remarkable that the pairs  $T_P3 - T_E1$  and  $T_P1 - T_E3$  show a higher  $\mathbb{J}\mathbb{M}\mathbb{I}(T_E, T_P; L)$ -value than the second best associated pair, although, the two columns do not show any dependence of each other at all. However, both distributions highly depend on the label vector, resulting in the high value calculated by this quantity. This in turn implies that a dependence between the two TFBSs is not required for a high value of this quantity as long as one or both distributions show any commonality to the label vector.

Using the  $\mathbb{C}\mathbb{M}\mathbb{I}$  results in  $\mathbb{C}\mathbb{M}\mathbb{I}(T_E1; T_P1|L) = 0$  for the best associated pair, indicating that  $T_E1$  and  $T_P1$  do not share any additional information about each other, if the label is known.

**Table 5.13.: Results of the synthetic generated count matrices.** Shown are the result values for all TFBS pairs for the joint mutual information ( $\mathbb{JMI}$ ), multivariate mutual information ( $\mathbb{MMI}$ ), conditional mutual information ( $\mathbb{CMI}$ ), dual total correlation ( $\mathbb{DTC}$ ) and pairwise mutual information ( $\mathbb{I}$ ). While the first four metrics consider the TFBS distributions of  $T_E$  and  $T_P$  as well as the label  $L$  of the PEIs (input or background PEI), the pairwise mutual information just focuses on the TFBS distributions in the input sequences, neglecting the generated background. All values are normalized by the alphabet size.

TFBS enhancer	TFBS promoter	$\mathbb{JMI}(T_E, T_P; L)$	$\mathbb{MMI}(T_E; T_P; L)$	$\mathbb{CMI}(T_E; T_P   L)$	$\mathbb{DTC}(T_E, T_P, L)$	$\mathbb{I}(T_E; T_P)$
$T_{P1}$	$T_{E1}$	<b>1.0</b>	<b>1.0</b>	0.0	<b>1.0</b>	0.0
$T_{P2}$	$T_{E1}$	0.43	0.43	0.0	0.43	0.0
$T_{P3}$	$T_{E1}$	0.5	0.0	0.0	0.5	0.0
$T_{P1}$	$T_{E2}$	0.43	0.43	0.0	0.43	0.0
$T_{P2}$	$T_{E2}$	0.43	0.43	0.43	0.86	0.0
$T_{P3}$	$T_{E2}$	0.43	0.0	0.43	0.86	0.0
$T_{P1}$	$T_{E3}$	0.5	0.0	0.0	0.5	0.0
$T_{P2}$	$T_{E3}$	0.43	0.0	0.43	0.86	0.0
$T_{P3}$	$T_{E3}$	0.0	0.0	<b>1.0</b>	<b>1.0</b>	1.0



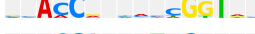



The non-associated pair gets the highest value by this quantity with  $\mathbb{CMI}(T_{E3}; T_{P3} | L) = 1$ . The second best associated pair results in a value of  $\mathbb{CMI}(T_{E2}; T_{P2} | L) = 0.43$  and thus, it is on position two in the ranking of the  $\mathbb{CMI}$ -values. Therefore, the  $\mathbb{CMI}$  predicts the pairs in the reverse order.

Using the  $\mathbb{DTC}$  results in  $\mathbb{DTC}$ -value of one for the best as well as for the non-associated pair. The second best associated pair gets a value of  $\mathbb{DTC}(T_{E2}, T_{P2}, L) = 0.86$  which implies that the  $\mathbb{DTC}$  does not reflect the order of the associated pairs.

Finally, I applied the pairwise mutual information ( $\mathbb{I}$ ) to the example dataset in order to demonstrate the importance of the background set and consequently, the requirement of the third variable  $L$ . For the calculation of the pairwise mutual information  $\mathbb{I}(T_E; T_P)$ , I considered only the input PEIs, since this quantity offers no differentiation between the input and the background. The  $\mathbb{I}$  results in  $\mathbb{I}(T_{E3}; T_{P3}) = 1.0$  for the non-associated pair and all other pairs have a value of 0, indicating that both binding sites do not offer any information about each other. This implies that the association between two TFBSs can in some cases only be captured by the consideration of the background set.

Summarizing my findings, the  $\mathbb{MMI}$  and  $\mathbb{JMI}$  arrange the pairs in the correct order according to their association strength.

**Table 5.14.: Inserted associated TFBSs in enhancer and promoter sequences with the representing logoplots.**

Pair	TFBS enhancer	TFBS promoter
1	V\$NFAT5_Q5_02 	V\$ROAZ_01 
2	V\$E2_01 	V\$GZF1_01 
3	V\$ZNF143_03 	V\$IRF1_01 

### 5.2.2. Simulation datasets

For a more conscientious evaluation of the different mutual information quantities, I constructed a collection of simulated sequence sets in which I artificially inserted associated TFBS pairs. In the first step, I trained two Markov chain models for the generation of synthetic enhancer and promoter sequences that show a nucleotide distribution close to natural sequences. For the generation of enhancer sequences, I trained the Markov chain model on p300 ChIP-Seq peaks provided by ENCODE (<https://www.encodeproject.org/>) of cell lines MCF7, IMR90 and K562 that do not have any overlap with known promoter regions. The model-generated enhancer sequences have an average length of 600bp and can vary in their length by +/-100bp. The Markov chain model for the generation of promoter sequences has been trained on a non-overlapping set of promoter sequences of genome wide RefSeq genes using the promoter region of -1000bp to +100bp relative to the transcription start site (TSS). The Markov model generated promoter sequences are of length 1100 bp. Using these models, I generated sequence sets each of which consisted of 1000 synthetic enhancer and 1000 promoter sequences and defined the  $i^{th}$  promoter sequence in the set to be paired to the  $i^{th}$  enhancer sequence.

In the next step, I inserted associated TFBSs in the sequences. In total, I chose three TFBS pairs (see Table 5.14) that are indicated by the names of the corresponding PWMs with an additional index for enhancer or promoter: V\$ROAZ\_01<sub>prom</sub> - V\$NFAT5\_Q5\_02<sub>enh</sub>, V\$GZF1\_01<sub>prom</sub>-V\$E2\_01<sub>enh</sub> and V\$IRF1\_01<sub>prom</sub>-V\$ZNF143\_03<sub>enh</sub>.

For each pair, I constructed a synthetic set of 1000 PEIs and only inserted one specified TFBS pair inside the sequences of a PEI set to avoid cross effects of different pairs.

In order to bring the synthetic example in line with realistic scenarios, I incorporated the “association strength” as an additional parameter. For this, I defined the association strength as the proportion of PEIs the inserted pair is important for, meaning, the fraction of promoter and enhancer sequences in which the two TFBS motifs show a dependence in their binding behaviour. Considering this, I assume that, in a set of PEIs, there is a certain kind of variability in the association strength of TFs on enhancer and promoter sequences. That means, some TF pairs are important for a huge number of PEIs of the entire set and therefore, the

TFBSs of enhancer and promoter sequences are strongly associated. However, some other TF pairs appear to be incorporated in a minority of PEIs and thus, the association strength of their binding site distributions is on a lower level. Addressing this point, I incorporated the association strength as a discrete variable with three states (*low*, *medium* and *high*) in my analysis where for a *low* association strength the pair is associated in 20% of all PEIs and for *medium* and *high* it is important for 50% and 90% of all input PEIs, respectively (see Table 5.15). To this end, I analyzed each of the three TFBS pairs for all three association strength resulting in total, in nine synthetic generated input sequence sets.

Motivated by the idea that the interplay between a certain TF in enhancer and a TF in promoter regions can be important for the PEI, although one or both considered TFs occur with a low frequency, I added the parameter “TFBS frequency” in the creation of the synthetic sets. Following this, I determined for each TFBS pair the number of single TFBS instances per sequence for the TFBS frequency states *low*, *medium* and *high*. Due to the fact that a fixed number of TFBS instances per sequence is unrealistic, I further incorporated a certain term of variability. Thus, the number of TFBS instances per sequence is determined by the TFBS frequency +/- the variability term whereas the TFBSs show a low variability in their binding site behaviour in the PEIs they are associated in and a larger variability in the PEIs, they are not associated (see Table 5.16).

To summarize, I have three different TFBS pairs each of which occurs with three different association strength and three different TFBS frequency levels resulting in 27 synthetic generated sequence sets under study each of which consists of 1000 PEIs.

**Table 5.15.: Visualization of the different states of the “association strength” variable.** The yellow and the blue TFs are important for the establishment of the underlying PEIs ( $\leftarrow\rightarrow$ ) or are not involved in the PEIs ( $\leftarrow\rightarrow$ ). Regarding the different association strength: for *low* the TF pair is important for/associated in 20% of all input PEIs, for *medium* in 50% and for *high* in 90% of all PEIs under study.

High	Medium	Low

**Comparison of the different multivariate mutual information metrics** For the analysis of the 27 generated synthetic datasets, I set the parameters for the overall workflow as

**Table 5.16.: Numbers of inserted TFBS instances for each artificially inserted associated TFBS pairing.** The numbers of insertions (# Motifs) varies according to the column “Variability” and further, the numbers differ among the PEIs the pair is important for/associated in (  $\leftarrow\rightarrow$  ) or not (  $\leftrightarrow$  ).

TFBS frequency	$\leftarrow\rightarrow$				$\leftrightarrow$			
	V\$ROAZ_01		V\$NFAT_Q5_01		V\$ROAZ_01		V\$NFAT_Q5_01	
	# Motifs	Variability	# Motifs	Variability	#Motifs	Variability	# Motifs	Variability
Low	1	1	0	0	2	0	2	1
Medium	3	2	1	0	5	1	4	1
High	5	3	2	2	7	1	6	1
	V\$GZF1_01		V\$E2_01		V\$GZF1_01		V\$E2_01	
	# Motifs	Variability	# Motifs	Variability	#Motifs	Variability	# Motifs	Variability
Low	1	1	0	0	1	0	2	0
Medium	2	1	1	1	3	0	3	0
High	3	2	2	1	7	2	6	1
	V\$IRF1_01		V\$ZNF143_03		V\$IRF1_01		V\$ZNF143_03	
	# Motifs	Variability	# Motifs	Variability	#Motifs	Variability	# Motifs	Variability
Low	3	2	2	1	4	0	3	0
Medium	3	2	2	1	7	2	6	1
High	3	3	2	1	9	3	8	2

follows: I filtered all columns in the matrix that had more than 50% of zero entries by setting  $t = 0.5$  and I set the number of intervals, the count values were assigned into, to  $q = 30$ . I further used a PWM-library of 166 matrices and ran the Match<sup>TM</sup>-algorithm by setting the parameter to *minimize the number of false positive* (minFP) predictions. After applying our approach to these datasets, I determined a TFBS pair to be significant if its mutual information value is  $\geq 0$ .

I applied all four different information theoretic quantities to these simulated sets and show the number of significant pairs of the MIII in Table 5.17. It can be seen that the number of significant pairs strongly depends on the association strength as well as on the TFBS frequency. Considering my findings of the determination of specific intra-regional cooperating TFs in the simulation dataset (see Section 5.1.2), the different numbers of significant pairs can be explained by the unintentional insertion of additional TFBS pairs that match to the inserted consensus sequences as well and their frequency of occurrence depends on the TFBS frequency and association strength constraints. The differences among the three pairs in turn can be attributed to the number of PWMs that match to the individual consensus sequences.



**Table 5.17.: Number of significant pairs identified by  $\mathbb{M}\mathbb{M}\mathbb{I}$  for the simulation dataset of each condition.**

	TFBS frequency	Association strength		
		Low	Medium	High
Pair 1	Low	3	5	12
	Medium	6	25	70
	High	34	56	106
Pair 2	Low	0	3	0
	Medium	5	6	9
	High	3	8	22
Pair 3	Low	3	5	12
	Medium	33	55	90
	High	46	65	124

Using a library of 166 PWMs, there are 27556 possible TFBS pairs between enhancer and promoters. Table 5.18 depicts the position of the inserted pair in the ranking of the different mutual information measures. For example, considering *Pair 1* with a *low* TFBS frequency and a *low* association strength, the  $\mathbb{M}\mathbb{M}\mathbb{I}$  votes *Pair 1* on rank one indicating that it shows the highest  $\mathbb{M}\mathbb{M}\mathbb{I}$ -value among all other potential TFBS pairs.

**Application of  $\mathbb{D}\mathbb{T}\mathbb{C}$**  Considering the performance of  $\mathbb{D}\mathbb{T}\mathbb{C}$  for *Pair 1*, it is high ranked in six cases. For a *low* or *medium* TFBS frequency with a *low* association strength, it was not identified at all. For a *high* association strength in combination with a *high* TFBS frequency, it is on ranking position 13. In the analysis of *Pair 2*, the  $\mathbb{D}\mathbb{T}\mathbb{C}$  correctly high ranks it in all cases except for a *low* TFBS frequency and a *low* association strength, for which the pair was not identified. For *Pair 3*, it is on top position in all cases.

**Application of  $\mathbb{C}\mathbb{M}\mathbb{I}$**  For the  $\mathbb{C}\mathbb{M}\mathbb{I}$ , the performance regarding the dataset of *Pair 1* is quite diverse and successful for *low* and *medium* TFBS frequency combined with *medium* and *high* association strength as well as *high* TFBS frequency and *low* and *medium* association strength. *Pair 2* has successfully been identified as important in all cases except for a *low* association strength combined with a *low* and *medium* TFBS frequency and for a *high* association strength and a *medium* TFBS frequency. Regarding *Pair 3*, it is on top in the pair ranking for all combinations regarding a *medium* and *high* TFBS frequency.

**Table 5.18.: Results for the simulation dataset.** The table gives the position of the inserted pair in the ranking of the underlying metrics for all condition combinations. In total, there are 27556 TFBS pairings participating in each ranking. Further, the number of intervals was set to  $q = 30$  and the threshold for zero entries filtering was set to  $t = 0.5$ .

	<b>TFBS frequency</b>	<b>Association strength</b>	<b>Rank</b> $\mathbb{DTC}(t_E, t_P, t_L)$	<b>Rank</b> $\mathbb{CMI}(t_E; t_P   t_L)$	<b>Rank</b> $\mathbb{JMI}(t_E, t_P; t_L)$	<b>Rank</b> $\mathbb{MMI}(t_E; t_P; t_L)$
Pair 1	Low	Low	-	-	-	-
	Low	Medium	1	1	1	2
	Low	High	1	1	1	1
	Medium	Low	-	-	-	-
	Medium	Medium	1	1	1	1
	Medium	High	1	11	4	1
	High	Low	1	14	17	1
	High	Medium	1	1	14	1
	High	High	13	222	24	1
Pair 2	Low	Low	-	-	-	-
	Low	Medium	1	1	1	6840
	Low	High	1	1	1	6745
	Medium	Low	1	324	1	1
	Medium	Medium	1	1	1	1
	Medium	High	1	423	1	1
	High	Low	1	1	2	1
	High	Medium	1	1	9	1
	High	High	2	2	80	1
Pair 3	Low	Low	1	767	1	1
	Low	Medium	1	533	1	1
	Low	High	1	1588	1	1
	Medium	Low	1	1	1	1
	Medium	Medium	1	1	1	1
	Medium	High	1	3	47	1
	High	Low	1	1	1	1
	High	Medium	1	1	1	1
	High	High	1	1	56	1

**Application of JMI** The JMI in general shows a mixed performance. *Pair 1* was successfully identified as most important in three cases: *low* TFBS frequency with *medium* and *high* association strength and *medium* TFBS frequency with *medium* association strength. *Pair 2* was high ranked for all combinations regarding a *low* and *medium* TFBS frequency except *low* TFBS frequency with *low* association strength. The JMI shows its best performance in the analysis of *Pair 3*, where it identifies it correctly on the first position for all combinations except *medium* TFBS frequency and *high* association strength and *high* TFBS frequency with *high* association strength.

**Application of MMI** The MMI identifies *Pair 1* as top candidate pair for all combinations of TFBS frequency and association strength except *low* TFBS frequency with *low* association strength as well as *medium* TFBS frequency combined with *low* association strength. Considering *Pair 2* for the condition of *low* TFBS frequency my approach was not able to identify it as the most important one but identified it correctly in all other cases. Regarding *Pair 3*, my approach successfully high-ranked the inserted pair throughout all conditions.

To summarize, the DTC and the MMI show the best performance in the analysis of the simulation datasets. However, regarding the first small example the performance of the DTC was inscrutable to some degree, since it is a combination of CMI and JMI. Therefore, I decided to use the MMI as the preferred metric for the determination of associated TFBSs between enhancer and promoter regions.

### 5.2.3. Comparison with MotifHyades

In order to compare the performance of my approach with an existing method, I applied MotifHyades([15]) to the simulation datasets. MotifHyades is a probabilistic approach published by Wong et al. in 2017, for the identification of *de novo* DNA motif pairs of paired sequences that is based on expectation maximization. In its first step the algorithm identifies DNA motifs using MEME and quantifies the discovery accuracy of motif pairs afterwards with two performance metrics [15]. For details please have a look at [15]. As input parameter, MotifHyades requests the number of pairs to detect and thus, the user needs to know beforehand how many pairs are likely representative in the paired sequence set under study. In order to increase the prediction probability for the inserted pair, I set the parameter *number of predicted pairs* =2. Regarding the results of MotifHyades (see Table 5.19), it performed in general well in most cases for a *high* TFBS frequency and a *high* association strength. In detail, MotifHyades predicted *Pair 1* in all cases (except *low* TFBS frequency and *low* association strength). However, the performance of MotifHyades dramatically dropped for the prediction of *Pair 2*, which has only been predicted for a *medium* TFBS frequency and a *high* association strength as well as for a *high* TFBS frequency in combination with a *medium* or *high* association strength. *Pair 3* was identified for all cases of *medium* TFBS

frequency as well as for *low* TFBS frequency with *high* association strength and *high* TFBS frequency in combination with *medium* or *high* association strength.

Summarizing my findings, MotifHyades performs well in the prediction of motif pairs of enhancer and promoter sequences, if the motifs occur with a *high* TFBS frequency and usually if they have a *high* association strength.

A comparing view of MotifHyades to the MIMI results reveals that both methods are not able to predict *Pair 2* with a *low* TFBS frequency but perfectly predict the inserted pairs for *high* TFBS frequencies and in general *high* association strength. In contrast to MotifHyades, the predictions made by MIMI are more reliable regarding associated TFBS pairs that have a *low* or *medium* association strength and/or *low* or *medium* TFBS frequency.

**Table 5.19.: Results of MotifHyades [15] in comparison to my approach using the MIMI** for the simulation dataset of all combinations regarding the association strength and TFBS frequency. An ✓ indicates a positive prediction of the inserted pair by MotifHyades (✓) or the MIMI approach (✓), while a ✗ represents the fail of MotifHyades (✗) or the MIMI-approach (✗).

Association strength	TFBS frequency	Pair 1	Pair 2	Pair 3
Low	Low	✗✗	✗✗	✗✓
Low	Med	✓✓	✗✗	✓✓
Low	High	✓✓	✗✗	✗✓
Med	Low	✓✓	✗✓	✗✓
Med	Med	✓✓	✗✓	✓✓
Med	High	✓✓	✓✓	✓✓
High	Low	✓✓	✗✓	✓✓
High	Med	✓✓	✓✓	✓✓
High	High	✓✓	✓✓	✓✓

#### 5.2.4. Comparative analysis of six human cell lines

**Dataset** I applied my method to six different human ENCODE cell lines: IMR90 (*fetal lung fibroblasts*), K562 (*leukemia mesoderm-lineage cells*), GM12878 (*lymphoblastoid cells*), HUVEC (*umbilical vein endothelial cells*), NHEK (*epidermal keratinocytes*) and HeLaS3 (*cervical cancer ectoderm-lineage cells*). Enhancer and promoter regions as well as their related PEIs for each cell line were taken from Whalen et al. [73]. In their study, Whalen et al. identified active enhancers and promoters using segmentation-based annotations, Roadmap Epigenomics and expression data from ENCODE (<https://>

(<http://www.encodeproject.org/>). The interactions between enhancers and promoters have been detected by Hi-C experiments [73, 74, 75].

**Table 5.20.: Number of enhancers, promoters and PEIs for the different cell lines.** The numbers reveal that enhancer as well as promoters can participate in more than one PEI (n:m relation).

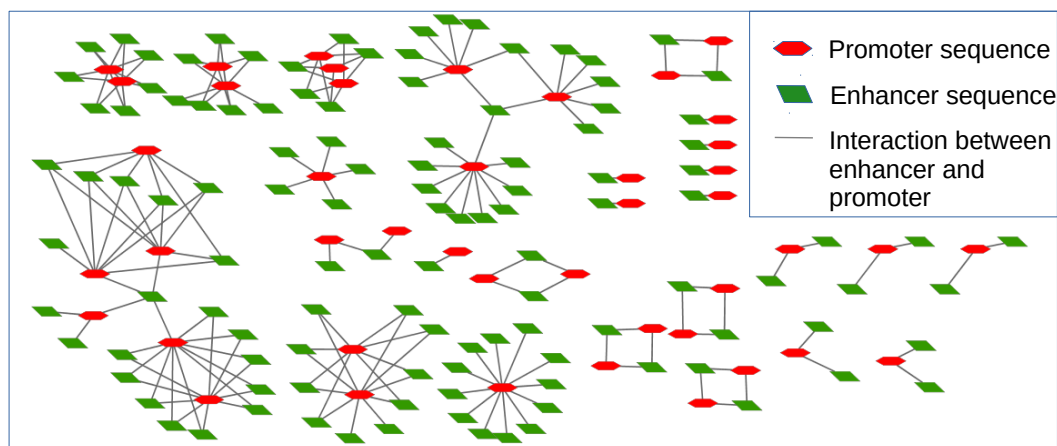
	<b>GM12878</b>	<b>HeLaS3</b>	<b>HUVEC</b>	<b>IMR90</b>	<b>K562</b>	<b>NHEK</b>
Enhancer	1932	1607	1390	1212	1742	1217
Promoter	736	474	562	422	619	304
PEIs	2113	1740	1524	1254	1977	1291

The numbers of enhancers, promoters and their interactions for each cell line are given in Table 5.20, where the number of enhancer sequences is about three times larger than that of the corresponding promoter sequences. In turn, the number of PEIs in a tissue is slightly larger than the number of enhancers, indicating that one promoter is paired with several enhancers and one enhancer is usually paired with one promoter but can also be paired to several promoter regions. The complexity of this  $n:m$  relation between enhancer and promoter regarding their pairing behaviour is exemplarily illustrated in Figure 5.10 as a network where red nodes correspond to promoters and green nodes to enhancer regions. The network consists of several unlinked sub-networks of different sizes some of them representing exclusive pairs of one enhancer and one promoter while some others form clusters of several enhancers and one promoter as centering node. Further, there are some larger sub-networks where additionally some enhancers are linked to several promoters. Independent of the tissue, the enhancer sequences are on average several hundreds base pairs in length. In turn the promoter sequences have on average a length of one to three thousand base pairs. Only IMR90 and NHEK have promoter sequences with an average length around 500 bps (see Table 5.21).

**Table 5.21.: Average length of promoter and enhancer sequences for each cell line.**

	<b>Average sequence length (in bp)</b>					
	GM12878	HeLaS3	HUVEC	IMR90	K562	NHEK
Enhancer	551	473	883	414	369	432
Promoter	2961	1411	2110	451	2380	511

The sequence length distribution of enhancer and promoter sequences is exemplarily shown in Figure 5.11 for cell line K562. Most of the enhancer sequences are of short length and



**Figure 5.10.: PEI sub-network, exemplarily taken from the K562 PEI-network.** This extraction consists of 160 nodes (116 enhancer and 44 promoter nodes) and 183 edges referring to PEIs and visualizes the  $n : m$  relation between enhancer and promoter regions.

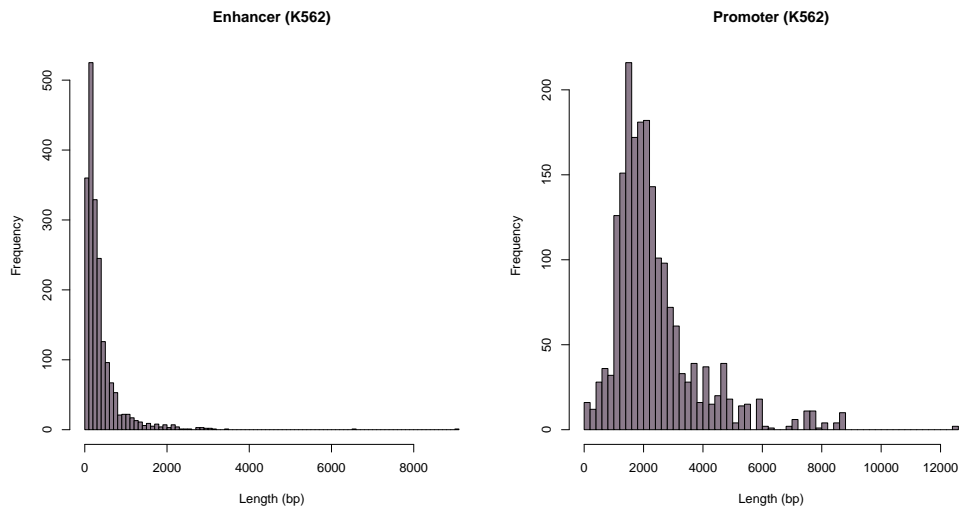
only a few outliers exhibit a similar length as most of the promoters. The length distribution of promoter sequences in contrast resembles a poisson distribution with maximum at about 2000 bps in length.

**MMII results** I applied my method for the detection of associated TFBS pairs in enhancer and their related promoter sequences based on multivariate mutual information (MMII) to the six cell lines and determined a TFBS pair to be significant if its MMII-value is positive.

**Table 5.22.: Summary of the identified inter-regional TFBS pairs using MMII for the different cell lines.** Shown are the number of TFBS pairs as well as the numbers of unique TFBSs of enhancer and promoter regions that are involved in the predicted TFBS pairs.

	GM12878	HeLaS3	HUVEC	IMR90	K562	NHEK
TFBSs promoter	21	19	59	1	38	2
TFBSs enhancer	19	16	40	1	19	2
TFBS pairs	53	39	217	1	95	2

The number of significant TFBS pairs ranges from one (IMR90) to 217 (HUVEC) significant TFBS pairs (see Table 5.22).



**Figure 5.11.: Length distribution of enhancer and promoter sequences for K562 cell line.**

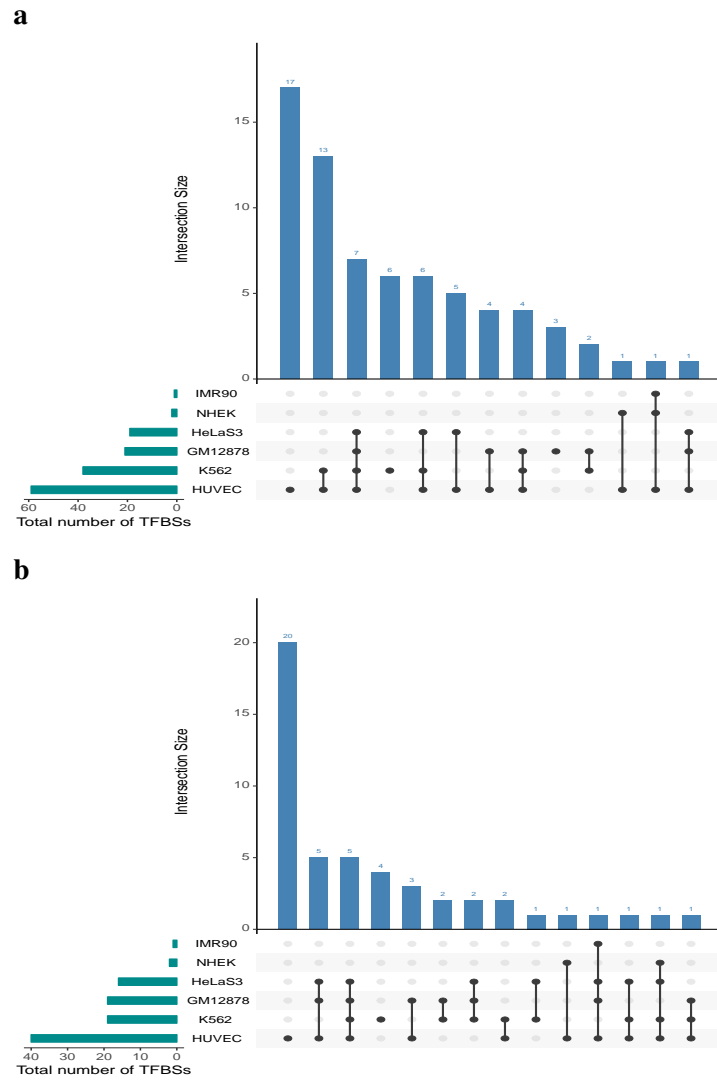
Figure 5.12 shows the number of unique and overlapping TFBSs participating in pairs for enhancer as well as promoter sequences.

Regarding the TFBSs of promoter sequences: 17 unique TFBSs were predicted for cell line HUVEC, six for the cell line K562 and three unique TFBSs for GM12878. There are seven TFBSs that appear to be important in the promoter sequences of cell lines HeLaS3, GM12878, K562 and HUVEC. However, there is no pair identified as important in all cell lines (see Figure 5.12).

Regarding the TFBSs of enhancer sequences, there are 20 unique TFBSs for cell line HUVEC and four unique TFBSs for cell line K562, whereas the other cell lines do not show any unique TFBSs. Five TFBSs overlap between HeLaS3, GM12878, K562 and HUVEC, but there are no overlapping TFBSs between all cell lines (see Figure 5.12).

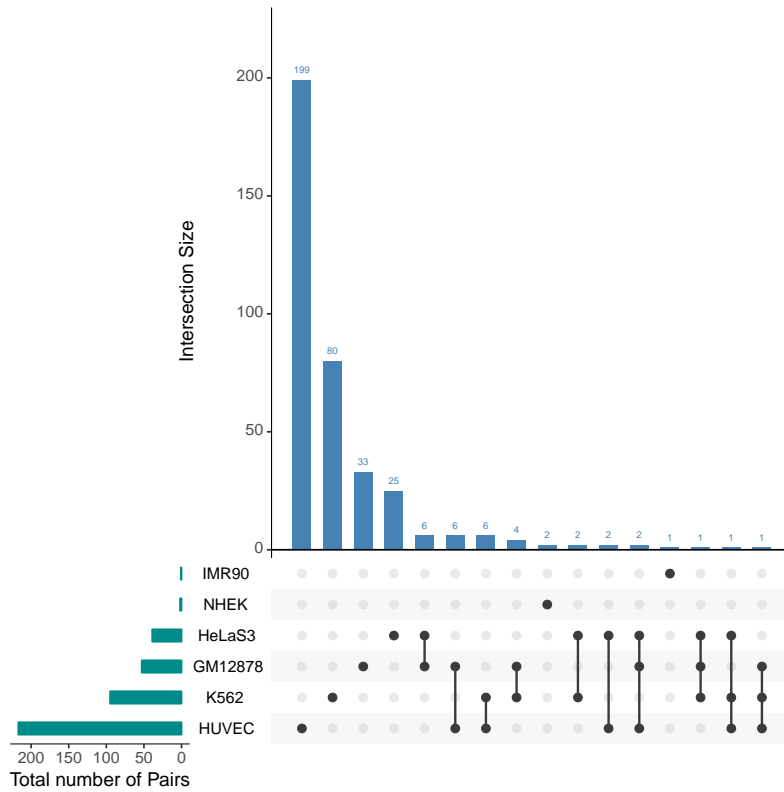
In contrast to the single TFBSs that build up the pairs, the distribution of overlapping and unique TFBS pairs looks rather different (see Figure 5.13). Most of the pairs are unique for a specific cell line, i.e. 199 pairs are unique for cell line HUVEC and all the pairs determined as significant in IMR90 and NHEK are unique for that cell line. The largest number of overlapping pairs is between GM12878 and HeLaS2 with six joint TFBS pairs. However, there is no pair that has been predicted as significant in all cell lines (see Figure 5.13).

For each cell line, the number of single TFBSs participating in pairs is smaller than the number of pairs itself. A closer look at the TFBS pairs itself reveals that some TFBSs are



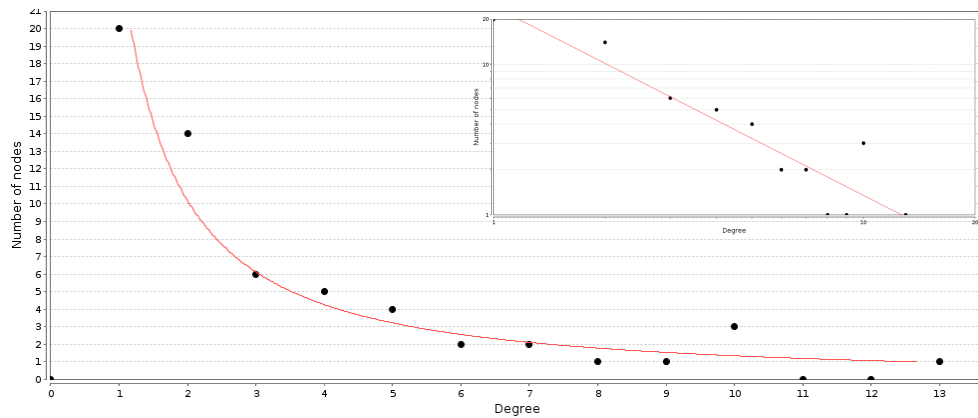
**Figure 5.12.: Number of unique and overlapping single TFBSs participating in significant pairs of the different cell lines for a) promoter and b) enhancer sequences, represented in matrix layouts using UpSet technique [72]. In the matrix layout, dark circles indicate the tissues that are part of the intersection.**





**Figure 5.13.: Number of unique and overlapping significant TFBS pairs of the different cell lines, represented in matrix layout using UpSet technique [72]. In the matrix layout, dark circles indicate the tissues that are part of the intersection.**

involved in a multitude of different pairs while some others are involved in one or a few TFBS pairings, indicating that the resulting cooperation network is scale-free (see Figure 5.14). Table 5.23 shows for each cell line the highly associated TFBSs participating in many pairs. It can be seen that most of these TFBSs are specific for the corresponding cell line and there is only a small overlap among the highly associated TFBSs in promoter regions.



**Figure 5.14.: Degree distribution of nodes of the K562 cooperation network** presented in linear and logarithmic (small plot) scale. The degree distribution can be fit to a power law distribution (red line) indicating that the network is scale-free.

These findings indicate that the differentiation between the lines is more difficult on the level of single TFBSs but quite obvious on the pair level.

**Table 5.23.: Highly associated TFBSs of the identified inter-regional TFBS pairs for the different cell lines.** Repeated occurring TFBSs are highlighted by background color. (For cell lines IMR90 and NHEC the determination of hub nodes is not possible due to its negligible small number of pairs.)

	GM12878	HeLaS3	HUVEC	IMR90	K562	NHEK
<b>Enhancer</b>	V\$IPF1_Q5	V\$TTF1_Q5_01	V\$MAF_Q6_01	-	V\$YY1_Q6_03	-
	V\$RFX1_01	V\$MAFA_Q4	V\$EBOX_Q6_01	-	V\$CREBP1_01	-
<b>Promoter</b>	V\$LUN1_01	V\$MEF2A_Q6	V\$MRF2_01	-	V\$LUN1_01	-
	V\$MEF2A_Q6	V\$SING4_01	V\$BBX_03	-	V\$SERALPHA_01	-
			V\$MEF2_03		V\$HNF1_Q6_01	
			V\$SREBP_Q6		V\$MEF2_03	

Having a more detailed look at the highly associated TFBSs reveals that V\$LUN1\_01,

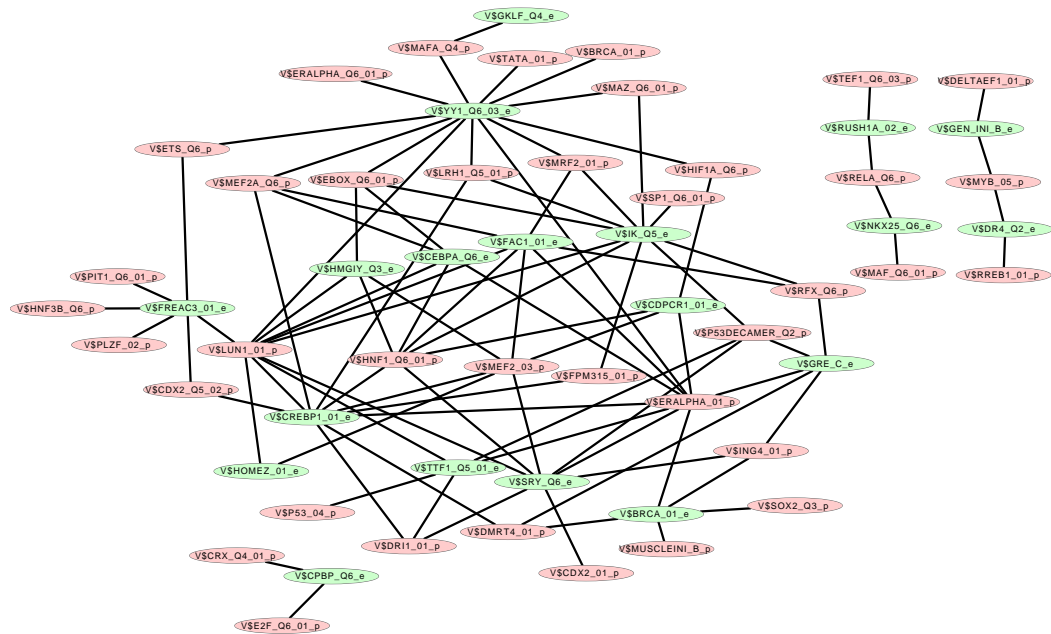
bound by *topoisomerase I binding arginine-serine-rich SUMO ligase* (TOPORS), is highly important for promoter binding in cell line GM12878 (*lymphoblastoid cells*) and cell line K562 (*leukemia mesoderm-lineage cells*). TOPORS mediates protein ubiquitination, is involved in cell cycle regulation, inhibits cell proliferation [76, 77, 78, 79, 80] and is associated with promyelocytic leukemia [76]. Further, V\$MEF2\_03 bound by the *myocyte enhancer factor 2A* (MEF2A), appears to play an important role for GM12878 and K562 and is known to be involved in mitochondrial organization, cardiac myofibril assembly as well as synaptic plasticity [81, 82, 83, 84, 85, 86, 87] and skeletal muscle differentiation [88]. The factor belongs to the MEF2-family that is in general known to be important for differentiation and morphogenesis [88]. Target genes of MEF2A are enriched in cell lines GM12878 and K562 [88, 89].

**Biological evaluation of K562 significant TFBS pairs** I chose cell line K562 to conduct a more detailed biological evaluation of my results. The cell line has been derived in the 1970s from a female patient with chronic myelogenous leukemia (CML) [90]. For this cell line I depict the TFBS association network in Figure 5.15 by enabling a differentiation between enhancer and promoter related TFBSs by color.

Regarding the enhancer TFBSs of K562 significant pairs, the most frequently TFBS was V\$YY1\_Q6\_03 that is bound by factor *Yin Yang 1* (YY1). This factor is known to be involved in the regulation of Notch-signaling as well as in the transition of G2-M phase in the cell cycle. It is further linked to adipogenesis, B-cell differentiation and neutrophil apoptosis [91, 92, 93, 94]. YY1 has been identified as an oncogene in a multitude of cancers and is over-expressed in acute and CML [95, 96, 97]. YY1 has been detected to contribute to structural interactions between promoter and enhancer regions in a similar way to CTCF protein [98] and in order to fulfill its regulatory functionality, it is known to bind to enhancer regions [95, 98] as well as to super enhancers [99]. Another conspicuous enhancer binding site motif is V\$CREBP1\_01 that is bound by the *activation factor 2* (ATF2). ATF2 is an histone acetyltransferase which is acting in calcium-mediated signaling, DNA repair and immune response [100, 101, 102, 103, 104, 105]. ATF2 has been detected to upregulate Fas/FasL in CML and in turn, the over-expression of Fas/FasL has been identified as a molecular commonality of these tumor cells [106]. The binding of ATF2 to enhancer elements has for example been detected at the interferon- $\beta$  enhancer [107]. V\$FAC1\_01 bound by *bromodomain PHD finger transcription factor* (BPTF or FAC1) participates in total in seven significant pairs of cell line K562, among which three pairs belong to the top ten pairs (see Table 5.24). FAC1 acts as a nucleosome dependent ATPase that stimulates cell proliferation and acts as a chromatin remodeling enzyme [108, 109, 110, 111, 112, 113, 114]. The most prominent function of BPTF is that it loosens the chromatin structure and thus, enables the DNA accessibility for other proteins. In this way, it is involved in the maintenance and differentiation of mammary gland stem cells, melanocytes and T-cells. Mutations of BPTF are associated with less accessibility of enhancer and promoter regions of genes that

are involved in the maintenance of adult hematopoietic stem/progenitor cells and in the activation of gene regulatory programs for hematopoietic stem cell functions [115]. In general, the chromatin structure plays an essential role in gene regulation and mutations in proteins involved in the remodeling of chromatin structure are often associated with different cancer types.

Regarding the promoter TFBSs, V\$HNF1\_Q6\_01 bound by HNF1A is highly represented. The factor is a transcriptional activator that functions in insulin secretion and fatty acid transport [116, 117, 118, 119, 120, 121].



**Figure 5.15.: TFBS association network between enhancer and promoter regions for cell line K562.** The nodes represent TFBSs predicted in enhancer (green) and promoter (red) regions. An edge represents the identified association between the binding site distributions of the underlying factors. For a further differentiation, the PWM names are extended with “\_e” for predicted in enhancer and “\_p” for predicted in promoter region.

V\$P53\_DECAMER\_Q2 bound by *tumor protein p53* (TP53) appears to be a highly associated TFBS in promoter sequences, since it is associated to four different enhancer TFBSs. TP53 is involved in regulatory processes of cell cycle arrest, apoptosis, senescence, DNA repair and keratinocyte differentiation [94, 122, 123, 124, 125, 126, 127, 128] and is incorporated in acute myeloid leukemia [129, 130] as well as adult acute lymphoblastic leukemia [131]. A mutation in TP53 gene was identified for cell line K562 [132, 133].

**Table 5.24.: Top ten associated TFBS pairs for cell line K562.** The first column gives the TFBS pair and the second column the number of PEIs the two TFBSs are simultaneously present.

<b>TFBS promoter</b>	<b>- TFBS enhancer</b>	<b>Number of PEIs</b>
V\$LUN1_01	- V\$FAC1_01	788
V\$MEF2A_Q6	- V\$FAC1_01	812
V\$MEF2A_Q6	- V\$YY1_Q6_03	720
V\$LUN1_01	- V\$IK_Q5	1390
V\$HNF1_Q6_01	- V\$CREBP1_01	947
V\$HNF1_Q6_01	- V\$FAC1_01	924
V\$MAZ_Q6_01	- V\$IK_Q5	839
V\$LUN1_01	- V\$TTF1_Q5_01	772
V\$LUN1_01	- V\$YY1_Q6_03	683
V\$ERALPHA_01	- V\$FAC1_01	910

Finally, my findings indicate that many of the TFBSs in enhancer and promoter regions are related to transcription factors that are in some way involved in cancer in general or in leukemia whereas some have already been described to be involved in CML.



### 5.3. Identification of inter- and intra-regional cooperating TFs in the context of inflammatory response in lung tissue

In this section, I applied the pointwise mutual information approach for the identification of intra-cooperating TFs as well as the multivariate mutual information based approach for the identification of intra-regional TF-cooperations to the same dataset, in order to demonstrate the mutual complementarity of the two methods. For this aim I chose a data set provided by the ExITox project. The ExITox project (FKZ 031L0120C) investigates the molecular changes in lung tissue in response to the inhalation of toxic substances (see <http://genexplain.com/exitox-ii/> for details). The underlying data set comprises 36 differentially expressed genes (DEGs) in response to butanol exposure.

For each gene, I took the promoter region -1000bp to +100bp relative to the TSS as promoter sequence under study. Further, I determined the potentially regulating enhancer regions for each gene by taking all enhancer regions provided by ENCODE that have a distance of at most 2 Mbp up- or downstream from the TSS of the gene and which have a length of at least 300bps. In total, I identified 1036 enhancer regions that take part in 2212 promoter-enhancer interactions (PEIs).

I applied my approach for the identification of intra-regional sequence-set specific TF cooperations to the enhancer sequences and to the set of promoter sequences, respectively. Further, I determined all potential inter-regional TFBS associations in the entire PEI sequence-set using the second approach. Finally, I end up with three TF cooperation networks, each for one analysis and summarized the networks in Table 5.25.

**Table 5.25.: Summary of the cooperation networks based on the intra- and inter-regional analyses.** The edges refer to identified cooperations and the nodes to the TFBSs. For the inter-regional cooperation network, I further distinguished between TFBSs in enhancer (enh.) and promoter (prom.) sequences.

	Intra-regional		Inter-regional
	Promoters	Enhancers	PEIs
<b>Edges</b>	36	44	170
<b>Nodes</b>	30	25	126 (51 enh. and 75 prom.)

In order to evaluate the performance of the approaches in the biological perspective, I determined the hub nodes for each TF cooperation network (see Table 5.26) and analyzed them according to their biological function by paying special attention to inflammatory processes in the lung.

**Table 5.26.: Hub nodes for the inter-and intra regional cooperating TFBS network.** The identified inter-regional hub nodes stem from the same network but are classified in enhancer and promoter TFBSs. The intra-regional hubs are taken from the network of enhancer and promoter sequence analysis, respectively.

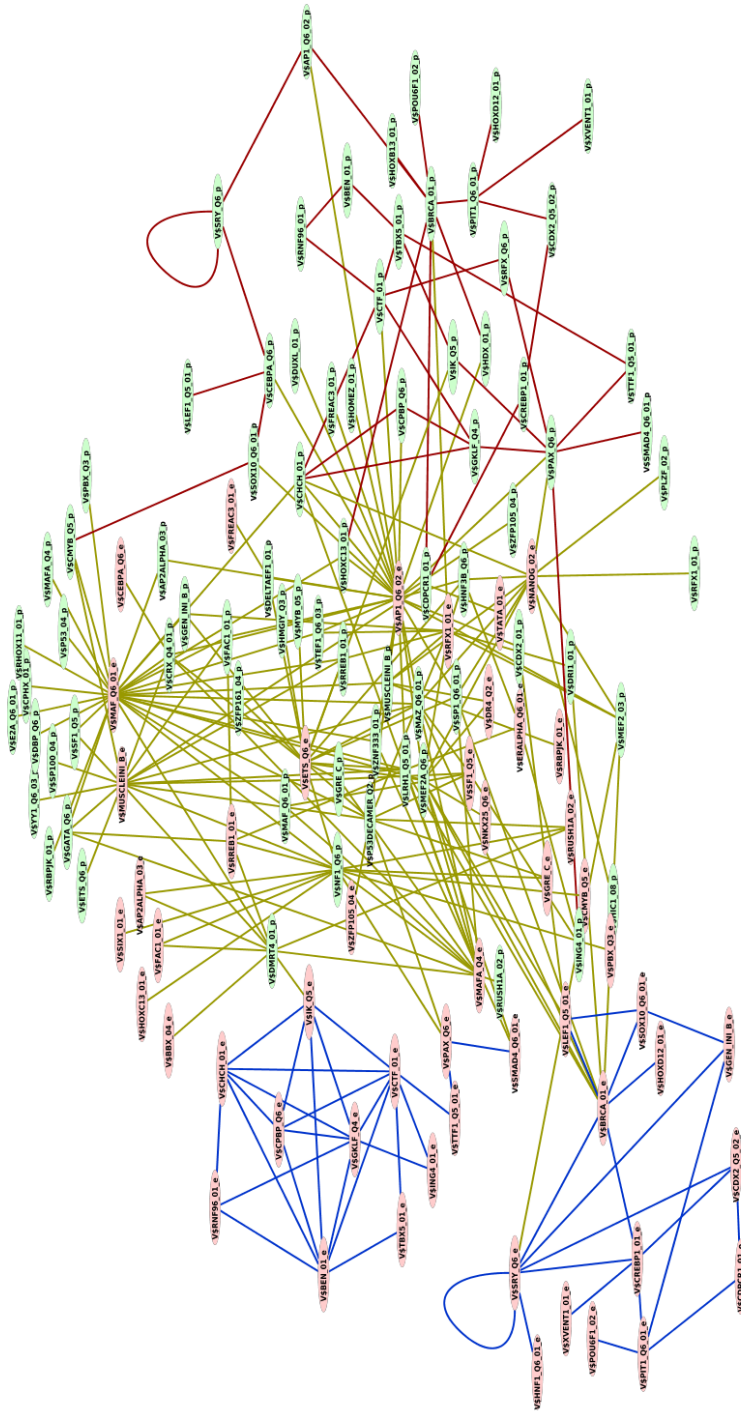
Intra-regional		Inter-regional	
TFBS promoter	TFBS enhancer	TFBS promoter	TFBS enhancer
V\$BRCA_01	V\$CTF_01	V\$NF1_Q6	V\$API_Q6_02
V\$PAX_Q6	V\$GKLF_Q4	V\$MEF2A_Q6	V\$MAF_Q6_01
V\$CTF_01	V\$BEN_01	V\$P53DECAMER_Q2	V\$MUSCLEINI_B
	V\$SRY_Q6	V\$DMRT4_01	V\$SETS_Q6
	V\$CHCH_01		V\$MAFA_Q4

In the analysis for the identification of intra-regional TF cooperations specific for the promoter sequences, the most frequently represented TFBSs are V\$BRCA\_01, V\$PAX\_Q6 and V\$CTF\_01. V\$BRCA\_01 is bound by transcription factor BRCA1 that is involved in apoptosis and heat shock response [134, 135]. V\$PAX\_Q6 is bound by paired box transcription factors such as PAX1 and PAX5 that are both involved in the immune system [136, 137] and PAX2 which acts in cell proliferation and antiapoptosis [138, 139, 140, 141, 142]. V\$CTF\_01 is bound by factors of the SMAD-family whose members act in response to TGF- $\beta$ , a cytokine, involved in fibrotic processes [143].

Regarding the identified intra-regional enhancer sequence-set specific TF cooperation network, one highly connected node is V\$GKLF\_Q4, which is bound by Krüppel-like factor 4 (KLF4), a factor that induces inflammation and apoptosis [144, 145, 146, 147, 148, 149, 150] and has been identified to attenuate lung fibrosis [151].

For the network of inter-regional TF cooperations, I differentiated between hub nodes that are related to TFBSs in enhancer and those in promoter sequences. Regarding the TFBSs in the promoter regions, V\$NF1\_Q6 is highly connected. V\$NF1\_Q6 is bound by factors such as NFIA and NFIB. NFIA acts in Notch signalling [152] and is linked to asthma plus rhinitis phenotype [153], whereas NFIB is involved in small cell lung cancer [154]. Another important binding site is V\$MEF2A\_Q6 bound by MEF2A, a factor upregulated in small-cell lung carcinoma [155]. The related factor MEF2D, which has a nearly identical DNA-binding domain and therefore an identical or very similar DNA-binding specificity,





**Figure 5.16.: Network of inter- and intra-regional cooperating TFs.** The network shows the aggregation of all three cooperation networks. The TFBSs related to promoter regions are colored red and those related to enhancer regions are marked in green. The edges refer to predicted inter-regional cooperations (yellow) and sequence-set specific intra-regional TF cooperations for the enhancer sequences (blue) and the promoter sequences (dark red).

has been identified to be upregulated in lung inflammation and the resulting development of lung cancer [155].

Regarding the highly connected TFBSs in the enhancer sequences, V\$AP1\_Q6\_02 is a highly connected node. V\$AP1\_Q6\_02 is bound by AP1 which might be activated by the development of oxidant/antioxidant imbalance in lung inflammation [156] and the inhibition of AP1 leads to the attenuation of lung inflammation [157]. Another highly connected binding site is V\$MAF\_Q6\_01 that is bound by MAF, a factor involved in toll-like receptor signaling and is also involved in immunity [158, 159, 160, 161, 162]. Further, the binding site V\$ETS\_Q6 is bound by ETS1 or ETS2. While ETS1 acts in apoptosis and cytokine secretion, ETS2 has been identified as a putative biomarker for progression of chronic obstructive pulmonary disease [163].

**Table 5.27.: TFBSs identified in the analysis for inter-regional and intra-regional TF cooperations.**

TFBSs promoter	TFBSs enhancer
V\$CHCH_01	V\$BRCA_01
V\$SOX10_Q6_01	V\$SMAD4_Q6_01
V\$HOXC13_01	V\$LEF1_Q5_01
V\$CPBP_Q6	V\$PAX_Q6
V\$BRCA_01	V\$IK_Q5
V\$CMYB_Q5	V\$SRY_Q6
V\$IK_Q5	
V\$PAX_Q6	
V\$CEBPA_Q6	
V\$HDX_01	
V\$AP1_Q6_02	
V\$ING4_01	
V\$CTF_01	
V\$CDPCR1_01	

There are several TFBSs involved in identified pairs of the inter- and intra-regional TF cooperation analysis. These TFBSs act as linking nodes between the inter- and intra-regional cooperation networks (see Figure 5.16 and Table 5.27). One of these linking TFBSs is V\$CPBP\_Q6 that is bound by KLF6, a factor involved in the activation of TGF- $\beta$  in the cellular response to the human respiratory syncytial virus [164]. Another linking TFBS is V\$CMYB\_Q5 bound by v-MYB which is among others involved in idiopathic pulmonary fibrosis [165]. Further, binding site V\$ING4\_01 is bound by ING4 which is involved in cell proliferation and apoptosis and in lung carcinomas [166, 167]. Finally, V\$CDPCR1\_01

bound by CUX1 appears to be a linking node between inter-and intra-regional TF cooperation networks. CUX1 acts in lung development, immune response and is downregulated in interstitial fibrosis [168, 169, 170, 171, 172, 173].

The complementary usage of the two approaches provides a more extensive insight in the underlying regulatory mechanisms in the cell, in contrast to the single analyses by joining the underlying TF cooperation networks. Thereby, TFs not conspicuous in the single analysis excel as linking nodes between the corresponding networks and, thus, can be identified as important factors in the regulatory processes of inflammatory response in the cell.



## 6. Discussion

In this chapter, I will discuss the methods established in this thesis as well as the corresponding results. First of all I will discuss the determination of intra-regional cooperating transcription factors based on the co-occurrence of their binding sites by using pointwise mutual information. Second, I will consider the identification of associated TFBSs between enhancers and their related promoters, referred to as inter-regional cooperations, by using and comparing different multivariate mutual information measures. In the last section of the chapter, I will discuss the complementary usage of these two applications based on the analyzed inflammation linked gene set.

### 6.1. Pointwise mutual information in the context of intra-regional cooperating TF identification

The pointwise mutual information ( $\mathbb{PMI}$ ) is an important measure in linguistics for the identification of word associations [41] as well as for document summarizing processes [42]. In their study, Bouma et al. [41] used the  $\mathbb{PMI}$  for the identification of word collocations in documents that share a certain kind of idiosyncrasy in their linguistic distribution. In turn, Aji S et al. [42] used  $\mathbb{PMI}$  for document summarizing processes. Thereby, they constructed a term-sentence matrix and identified important words for each sentence using  $\mathbb{PMI}$  under the consideration of the entire distributions of words and sentences in the document. Inspired by these two studies, I adopted the  $\mathbb{PMI}$  from the field of linguistics to the field of bioinformatics in order to identify collaborating TFs based on the co-occurrence of their binding sites as well as important single binding sites for a certain sequence in consideration of the entire sequence set. Thereby, I considered a sequence set as a document, sequences of this set as sentences and transcription factor binding sites (TFBSs) as words in these sentences.

In higher organisms, the interplay between TFs is usually more important for a proper gene regulation than the single factor itself. In order to collaborate with each other, the factors form non-random combinations of dimers or high order complexes and the underlying binding sites of the factors appear to be located next to each other on DNA. Thus, as confirmed in a multitude of studies [2, 3, 4, 5, 6, 7, 9, 10, 174], the distribution of TFBSs in a set of regulatory sequences offers information about which factors are cooperating with each other. Therefore, the aim of this study was to identify cooperating TFs based on their binding sites. However, the computational prediction of TFBSs suffers from false positive

predictions. Further, there are some TFBSs that are highly over-represented and can be considered as a kind of punctuation marks or stop words like "a", "the", "of". These words are important for the grammatical structure of the sentence but do not provide any information about the general meaning of the sentence. Some other TFBSs are highly underrepresented like nouns that occur just one or a few times in the whole text. The filtering of these highly over- or underrepresented TFBSs was a challenging task of this study and was carried out in *phase 1* and *phase 2* of the algorithm (see Section 4.1.1).

Further, some predicted binding sites of the same type tend to overlap with each other and can be interpreted as redundant words in the context of linguistics that do not provide any further information to a sentence under study. However, by considering all these binding sites despite their overlap would result in an overestimation of these sites. Therefore, the filtering of these overlapping factors is crucial to avoid this overestimation and was conducted in *phase 3* of the algorithm (see Section 4.1.1).

Since my approach deals with the recognition of significant co-occurring TFBSs, I have to define TFBS pairs according to their localization on DNA. A well accepted approach is the definition of pairs according to the distance of the binding sites. For this aim, two distance constraints are well accepted: i) the determination of the preferred distances [2, 6]; ii) predefined minimal and maximal distance thresholds [8, 13, 175]. In this study, I constructed pairs by using predefined minimal and maximal distances as suggested by Hu et al. [8]. However, Hu et al. determined the distance between two TFBSs as the difference between the last nucleotide of the first TFBS and the first nucleotide of the second TFBS [8]. I did not follow this distance definition, since I allowed a certain kind of overlap between binding sites that would result in negative distances. Further, the borders of TFBSs predicted by PWMs are fuzzy and a distance definition based on the borders is not convincing. Addressing these points, I defined the distance of two TFBSs as the distance of their centers.

I used the average product correction (APC) theorem for a further elimination of noise arising from false positive TFBS predictions. The APC-theorem was proposed by Dunn et al. [48] for the estimation of noise of residue positions in multiple sequence alignments based on information theory. Since the approach is universally applicable for similar data structures, I applied it to estimate the background  $\mathbb{PMII}_{pc}(t_a, t_b)$ -values for a pair of TFBSs  $t_a$  and  $t_b$  in consideration of the sequence set under study. Afterwards, I subtracted the background  $\mathbb{PMII}_{pc}(t_a, t_b)$ -values from the observed  $\mathbb{PMII}_{pc}$ -values in order to separate noise from signal arising from functional collaborations. The resulting signal  $\mathbb{PMII}_{pc}$ -values were later used for the following determination of significant pairs.

In order to demonstrate the performance of the method, I constructed a synthetic sequence set where I inserted a TFBS pair that has been successfully identified by my method. I further performed a comparison study of my approach with existing methods. It turned out that all methods identified different sets of TFBS pairs as important and showed only a small

number of overlapping pairs. This indicates the different biological and computational considerations and assumptions made, regarding the interaction of TFBSs in the development of the different methods. However, all methods showed a similar performance in the statistical evaluation. These findings are supported by Klepper et al. [176] in his comparison study of several TFBS pair detecting methods in which no method performed remarkably better than the other. In order to depict a broad spectrum of important TFBS pairings, I recommend to use all methods together.

For a biological evaluation of the new method, I applied it to a genome wide set as well as on a breast cancer gene set. In both analyses, the underlying transcription factor interactions of 44 significant TFBSs pairs out of all significant pairs were confirmed by experimental findings reported in literature. There are 10 and 20 pairs determined as significant in the whole genome as well as in the breast cancer gene set analysis that have not yet been experimentally validated. Three of these pairs (V\$CETS1P54\_01-V\$MYCMAX\_B, V\$CP2\_01-V\$SF1\_Q6 and V\$SOX9\_B1-V\$STAT6\_01) are significant in both analyses. As described in [177] a general reason for the significant co-occurrence of the unconfirmed pairs can be that they do not interact directly and physically, but indirectly through an additional co-factor. A further reason is the lack of experiments for these specific protein-protein interactions.

As suggested by Hu et al. [8], I performed the analysis using different distance constraints for the TFBS pair construction. The results revealed that there is still a certain kind of overlap between the significant pairs of different distance constraints, indicating the robustness and consistence of the results obtained by the method. Based on the significant pairs, I constructed collaboration networks for each input sequence set under study (see Figures 5.1 and 5.2) in order to explain the potential biological functions of the TFBS pairs and to explain the preferred binding behaviour of these factors. In agreement to Hu et al. [8] the collaboration networks split into two unconnected subgraphs, where one subgraph shares more AT-rich binding sites as GC-rich binding sites and vice versa. These findings indicate that the general collaboration network of TFs is split into two major groups based on their binding behaviour.

In order to overcome the influence of false positive predictions to some extent, I applied the average product correction (APC) theorem in my study for the determination of background co-occurrences resulting from false positive predictions. However, the results revealed that there is a strong overlap between the significant pairs of the individual sequence sets, indicating that the power and functionality of the APC theorem is insufficient to handle the remaining obstacles for the identification of sequence-set specific TF cooperations. Thus, I extended the original method in order to separate the significant pairs into sequence set specific and common/general important ones by creating background sequence sets based on shuffling the original sequences using uShuffle-algorithm[36]. Thereby, the general nucleotide composition as well as the core of TFBSs is maintained by setting the  $k$ -mers size

to  $k=3$ . The influence of parameter  $k$  is as follows: the level of subtracted background co-occurrences decreases/increases with  $k$ . Thus, enlarging  $k$  leads to an increased background level while the reduction of  $k$  is followed by a reduced background level.

The parameter  $\alpha$  is used for linear scaling of the subtracted background level and thus, to reduce or enlarge its effect on the original  $\text{PMII}_i$ -values. Setting  $\alpha = 0$  results in the subtraction of the  $\text{AVG}(\text{PMII}(t_a; t_b))$ -value itself, while for  $\alpha = 1$  the doubled  $\text{AVG}(\text{PMII}(t_a; t_b))$ -value is used for the identification of sequence set specific pairs. An  $\alpha = -1$  results in the original significant pair analysis without the determination of specific pairs. Although the parameter  $\alpha$  linearly influences the level of subtracted background co-occurrences, its effect on the number of significant pairs appeared not to be linear. However, the effect of  $\alpha$  strongly differed between the individual sequence sets and seemed to be sequence set specific.

For a biological evaluation, I performed a comparison study of the original significant pairs and the specific pairs identified in the extension approach for five breast cancer subtype related gene sets. Eight pairs have been determined to be significant throughout all breast cancer subtypes, whereas, in turn no specific pairs have been identified in the extended approach indicating that the extension successfully separates specific TFBS pairs from common ones. Further, the resulting collaboration networks changed their structure according to  $\alpha$ . Thereby, some hub nodes in the original collaboration networks kept their property of being a hub, some others lose a majority of their interacting partners and were afterwards only of low degree in the collaboration network of specific pairs whereas some other nodes became hubs in the new network.



## 6.2. Multivariate mutual information in the context of inter-regional cooperating TFs

The ability for the interaction of two gene regulatory regions, like enhancer and promoter regions, is formed by the transcription factors binding to these regions. Thereby, some transcription factors are more important for this interaction than others and stay in association with other factors on the pairing DNA region. Thereby, the binding behaviour and thus, the binding site distributions are associated with each other. I tried to measure the level of dependence using different mutual information metrics that consider three random variables. The third random variable provides information about the origin of the underlying data since I created a background sequence set. This background sequence set is required to decrease the effect of false positive TFBS predictions that can lead to false positive associations between two TFBSs. Thereby, noise is separated from the signal arising from real TFBS associations. The background sequences are created using uShuffle [36] algorithm

that keeps the general nucleotide composition as well as the frequency of  $k$ -mers in the input sequences. I evaluated the performance of the  $\text{MMI}$  for different  $k$ . It turned out that the performance for  $k \in \{1, 2\}$  appeared to be continuously of high quality. However, this can be explained by the fact that some binding sites did not occur in the background sequences and the pairs are thus high ranked due to the lack of their corresponding binding sites in the background sequences. In turn,  $k \in \{4, 5\}$  kept the sequences too similar to the original ones and no differences in the binding sites counts could be determined. Following these findings, I kept  $k = 3$  (as I did in the extended version of my first approach) enabling that TFBSs still occur in the background sequences but the count value distributions differ from those of the original input sequences if their binding sites have any biological importance.

In order to avoid the correlation of TFBSs due to zero count values on both sides, I filtered all TFBSs that have more than 50% zero entries in the input sequence set.

For the purpose of a proper comparison between the count value distributions of the TFBSs, I first normalized all count values and, second, assigned them to predefined intervals. As normalization strategy, I chose the *min-max normalization* using global minimum and global maximum count values. Using the column minimum/maximum led to a poisson distribution of the normalized count values in the range between zero and one, and thereby, complicating the differentiation between the individual distributions. Using the global minimum/maximum relocated the original count value distribution to the range between zero and one and kept the original distribution properties. After the normalization of count values, the count values were assigned to  $q + 1$  intervals where  $q$  intervals are equally distributed in  $[0, 1]$  and one additional interval. All values of zero are assigned to this additional interval in order to differentiate between a low number count and the non existence of this binding site in a sequence.

I performed a comparative study between four different mutual information measures regarding three random variables: dual total correlation ( $\mathbb{DTC}$ ), conditional mutual information ( $\mathbb{CMI}$ ), joint mutual information ( $\mathbb{JMI}$ ) and the multivariate mutual information ( $\mathbb{MMI}$ ). The definitions of the different measures are given in Chapter 3. For this, I generated synthetic paired sequence sets and inserted three TFBS pairs in these sets by constructing different conditions regarding the TFBS frequency and the association strength of each pair. Although the  $\mathbb{CMI}$  performed well in most cases for the synthetic sequence set, it performed poorly in the small starting example. It turned out that  $\mathbb{CMI}$  is not able to predict perfect associated pairs, if the information of the label does not offer additional information. The  $\mathbb{JMI}$  performed well in most cases for both sets. However, it high-ranks some pairs that do not show any association with each other, but one binding site distribution is somehow in dependence on the label distribution (see TFBS pair  $T_{E1} - T_{P3}$  in Table 5.9). The  $\mathbb{MMI}$  and  $\mathbb{DTC}$  clearly outperformed the other two measures. Although the  $\mathbb{DTC}$  consistently identified the inserted pair, a closer look to its predictions revealed that its results strongly depend on the distribution of just one TFBS binding site. Since the  $\mathbb{DTC}$  is build of the  $\mathbb{CMI}$  and the  $\mathbb{JMI}$  its prediction performance is not reliable and binding sites ranked high that do not show any association linked to the origin of the data (input and background set). Therefore, I stayed with the  $\mathbb{MMI}$  for further analysis of real biological data.

In order to evaluate my method, I compared its performance with MotifHyades [15], a tool published by Wong in 2017. It turned out that MotifHyades performed quite well on the synthetic sequence sets. However, for low numbers of TFBSs or low association strength of the TFBS pairs, MotifHyades performed poorly and the  $\mathbb{MMI}$ -approach clearly outperformed it. As explained by Hu, the algorithm has been developed for predicting statistically significant over-represented TFBS pairs of enhancer and promoter sequences. Thus, low associated pairs are not targeted.

I decided to go without a statistical analysis which would be based on data bases like BioGRID, TransCOMPEL or STRING, since this approach is not targeting the direct physical interaction between two transcription factors. It is much more likely that the cooperation is mediated by other factors such as co-factors. Exceeding the definition of true interactions in a way that direct as well as third-party interactions are included would lead to the fact that every factor can interact with every other factor throughout some highly connected factors such as EP300. Consequently, such a statistical analysis would not be meaningful.

I analyzed known promoter enhancer interactions based on ChIA-PET data of six human cell lines in order to determine the transcription factors that play important roles for the formation of these interactions. It turned out that the single TFBSs forming the identified pairs show a huge overlap between the different cell lines. In contrast, the overlap of the determined TFBS pairs themselves appears to be rather small suggesting that the differences in gene regulation are more on the level of paired transcription factor interactions than on the single factors. This finding is in consideration with that of my first approach where the overlap between the intra-regional TF cooperations was much higher than that of the single

binding sites themselves. Thus, both results support the hypothesis that single TFs and their binding sites are re-used for different purposes, e.g. cellular contexts and, consequently, a flexible and specific gene regulation is mainly based on the combinatorial binding of TFs.

As exemplarily shown for cell line K562, the degree distribution of the nodes follows a power-law distribution, and thus, the network is scale-free. This, in turn, reveals that some TFs participating in many pairings and are represented as hub nodes in the network while the majority of TFs is only involved in a few pairings. Consequently, some TFs are of major importance for the regulation of the underlying gene set but have to cooperate with other factors in order to fulfill their regulatory functions. These highly interacting TFs are presented as hubs in the underlying cooperation networks. The biological evaluation of these factors showed that some of them have already been linked to the analyzed cell lines or their corresponding phenotypes (i.e. leukemia).

Comparing the hub nodes of the different cell lines reveals that there are no overlapping hubs representing enhancer TFs. In turn, there are two transcription factors identified in promoter sequences the binding sites of which represent hub nodes in at least two cell lines: TOPORS and MEF2A. Members of the MEF2A family are known to be involved in the upregulation of genes in cell lines K562 and GM12878 [88, 89]. TOPORS in turn is known to be involved in promyelocytic leukemia [76].

A biological evaluation of the identified associated TFBS pairs for cell line K562 (a chronic myeloid leukemia (CML) cell line) points out that i.e. the identified transcription factors YY1 and ATF2 are both known to be involved in CML and are enhancer binding factors [95, 96, 97, 98, 106].

### 6.3. Complementarity of PMII and MIII in a biological application

I demonstrated the complementarity of the two approaches based on a differentially expressed gene set that is linked to the inflammatory response in lung tissue. For this data set, I applied the first approach in order to determine sequence-set specific intra-regional TF cooperations of enhancer and promoter sequences, respectively. I further identified inter-regional TF cooperations for the underlying promoter-enhancer interactions using the second approach.

The majority of TFs representing hub nodes of both, enhancer and promoter sequence-set specific intra-regional TF cooperation networks, are linked to inflammatory or fibrotic reactions in general. Some of them have already been described to be involved in inflammatory reactions in lung tissue such as KLF4 which is known to be involved in attenuate lung fibrosis [151].

Regarding the inter-regional TF cooperations, a lot of the TFs representing hub nodes in the underlying networks are linked to inflammation or fibrotic processes such as ETS2 which is a putative biomarker for progression of chronic obstructive pulmonary disease [163].

Combining the cooperation networks of both analyses provides new insights in the underlying regulatory mechanisms by uncovering the linking nodes of the networks that participate in inter- and intra-regional TF cooperations. These linking TFs are not identifiable by one of the single analysis and cannot be determined by single binding site enrichment analyses since they are not necessarily enriched in the sequence-set under study. Thus, the identification of these factors can only be achieved by the combination of both approaches and the biological evaluation of these factors confirmed their impact in inflammatory processes, i.e. v-MYB is involved in idiopathic pulmonary fibrosis [165].

## 6.4. Impact of combinatorics in transcription regulation

A transcriptional regulation network in general consists of nodes representing TF genes and edges that represent a regulatory relationship between them. In contrast to signal transduction, metabolic and protein-protein interaction networks, the topology of the transcriptional regulation network is not scale-free. This general finding reveals that the network displaying the relation of single TFs and their target genes resembles a random generated network (i.e. Erdős-Renyi network) and, thus, the binding of a single TF to a certain gene appears to be of low importance in general.

In my work, I generated intra- and inter-regional TF cooperation networks as well as joint networks consisting of both, intra-and inter-regional TF cooperations. In these networks nodes correspond to TFs and edges to predicted cooperations between them. It turned out that all these networks own the scale-free property as it is expected for protein-protein interaction networks and implies that the topological structure of my networks is not randomly constructed and some TFs are strongly linked to other TFs while the majority of TFs is sparsely interconnected.

For both approaches, I performed a comparison study between the analyzed cell lines regarding the overlap of single TFs participating in pairs and the pairs themselves. It turned out that the overlap between single TFs is much higher than that of TF pairs, which underpins the general assumption that the specificity of gene regulation is based on the combinatorial acting of TFs rather than on the single factor itself. Comparing the hub nodes of the individual TF cooperation network reveals TFs that are highly interconnected in the TF cooperation networks and seem to be key players in the regulation of the analyzed cell lines. The overlap of hub nodes among the different cell lines is rather small, supporting the hypothesis that the specificity of gene regulation of each tissue is based on some striking factors which cooperate with a multitude of other factors in order to fulfill their regulatory functions.

Although the overlap of the single TFBS involved in predicted pairs is huge, there are only a few TFBSs representing hub nodes in several constructed TF cooperation networks. This, in turn, supports the generally accepted assumption that the specificity of gene regulation is based on the pairing of TFs. This is well known for composite elements which form the smallest function unit within which protein-protein as well as protein-DNA interactions contribute to a highly specific transcriptional regulation [178]. These composite modules present a crosstalk between different regulatory pathways [178] which further underpins their specificity. Considering my findings that tissue specificity is reflected by TF pairs and not by single TFs, it is very likely that most of my predicted inter- and intra-regional pairs also present a crosstalk between different regulatory pathways, although they are not necessarily composite elements.

Eukaryotic TFBSs are in general relatively short and unspecific and, thus, can be bound by

a multitude of TFs. However, under a specific condition or in a certain cellular context, the binding site has to be bound by a defined factor in order to fulfill its regulatory function. The specific binding of this factor under the right condition is thereby not only in dependence on the binding site sequence but also on the protein neighbourhood of the factor formed by the already bound TFs. Thus, the complementarity of TFs in the 3D structure of neighboured proteins is of high importance and again, although the TFBSs themselves are rather unspecific, the combination of TFs in its surroundings enables a specific gene regulation.

It is widely accepted that the binding of a certain TF on DNA can enable the binding of other factors by bending the DNA helix or altering DNA structure in general. These aspects can be extended by my findings regarding the TFs that participate in inter- as well as intra-regional TF cooperations. The existence of these overlapping factors might lead to the suggestion that the factor of the enhancer is required to form the pairing between the two factors bound to the promoter. This can be based on the direct physical interaction in which the TF bound to the enhancer simply stabilizes the binding of one or two factors on the promoter e.g. by direct interaction or by a co-factor. It might also be possible that the factor bound to the enhancer leads to the modification or bending of the DNA structure of the promoter in a way that establishes the interaction between the two promoter bound factors through an indirect way without directly interacting with them. Another conceivable scenario is the formation of intra-regional TF cooperations which recruits co-factors that, in turn, interact with factors of enhancer regions and, thus, the inter-regional TF cooperation can only be established by the previous intra-regional TF pair formation.

All these findings are in line with the general knowledge that eukaryotic transcriptional regulation is based on the combinatorial binding and interacting of different TFs. However, I cannot capture the way the TFs cooperate with each other with my approaches and can only make assumptions, whether two TFs act in an agonistic or antagonistic manner or in a direct physical interaction or indirectly via co-factors or just in a functional manner.

## 7. Conclusion

In this last chapter, I first summarize the methods established in this thesis as well as their results and contributions. Afterwards, I give an outlook in which I provide some ideas for method extensions and list some potential fields of applications for future research interests.

### 7.1. Summary

In this thesis, I developed two different information theoretical approaches for the identification of potential cooperating transcription factors based on their predicted binding site distributions in a sequence set under study. In the first approach, I used the pointwise mutual information for the identification of potentially cooperating TFs inside a regulatory DNA region (intra-regional cooperations) based on the co-occurrence of their TFBSs (see Section 4.1). Since the pointwise mutual information is a powerful tool in linguistics for the determination of word collocations and document summarizing processes, I consider TFBSs as words and DNA sequences as sentences in a document. This method appeared to be very successful in comparison to existing methods and in the application to a synthetic gene set. However, the predicted TFBS pairs between different tissues are highly overlapping and targeting this point, I extended the original method in order to separate common (ubiquitously) occurring TF cooperations from sequence-set specific ones (see Section 4.1.2). Therefore, I created background sequence-sets that preserved the general nucleotide composition as well as the number of tri-nucleotides and thus, the core of TFBSs and used these sets to estimate for each TFBS pair the level of background co-occurrences. I applied the extended approach to gene sets of five breast cancer subtypes and successfully separated common TFBS pairs from sequence-set specific ones.

In my second approach, I developed a method based on mutual information for the identification of associated TFBSs between promoter and their related enhancer regions (see Section 4.2). In analogy to the extension of the first method, I created background sequence-sets by preserving the general nucleotide composition and the core of TFBSs, and directly integrate it in the calculation of mutual information by using a third random variable that indicates the origin of the data (i.e. input or background). I applied my approach with four different mutual information metrics to simulated data sets and compared their performance. I concluded that the multivariate mutual information (MMI) is most propitiate for my purposes and conducted an analysis of six human cell lines using MMI and performed a biological evaluation of these findings.

Further, the cooperation networks reveal TFs that participate in a multitude of pairings and are presented as hub nodes in the network. These TFs seem to play an important role in the transcriptional regulation of the underlying gene sets and might not have been identified in an enrichment analysis for single TFBSs since they are not necessarily enriched in the sequence-set under study. The combinatorial usage of both approaches resulted in a joint network of inter- and intra-regional TF cooperations which, in turn, offered TFs that act as linking nodes between the inter-regional and the intra-regional TF cooperation network. The biological evaluation of these factors pointed out their importance for the underlying molecular mechanisms. The identification of these nodes is only possible by the combination of the two approaches and cannot be achieved by e.g. a single TFBS enrichment analysis.

Finally, both methods are able to identify functional or physical cooperations between TFs that appear to play a critical role in the regulation of the gene set under study. These pairs provide new insights in the general understanding of transcriptional gene regulation and are new targets for laboratory experiments.

## 7.2. Outlook

Regarding the detection of intra-regional TF cooperations based on the co-occurrence of their binding sites it might be worthwhile to extend the approach for three or more co-occurring TFBSs in order to overcome the limits of pairwise identifications. Therefore, the pointwise mutual information measure needs to be extended to three or more random variables and a significance threshold needs to be established that is able to compare the resulting pointwise mutual information values of TFBS pairs to those of TFBS triplets or higher order complexes.

Although the identification of associated TFBS pairs of promoters and their related enhancer regions appears to be successful in the first place, the separation of the promoter-enhancer interactions (PEIs) into *the identified pair is directly involved in the PEI* and *the identified pair is not directly involved in the PEI* is not possible yet and needs to be established. In this context, it might also be worse to incorporate the locations of the binding sites in the analysis and consider only pairings that underlie certain distance or position constraints.

Both methods still suffer from the redundancy of the underlying PWMs that lead to a multitude of identified significant TFBS pairs that are based on the same sequence parts. Therefore, a method needs to be incorporated in the analysis for clustering all predicted pairs that stem from the same sequence regions. As a result, some other pairs will move up of which the TFBSs are not multiple times presented in the sequences but are still important for the gene regulation.

The identified TFs belonging to inter- and intra-regional cooperating TF pairs of both methods provide new insights in the complex process of gene regulation and can help to properly



identify the underlying cellular pathways and master regulators in combination with single TFBS enrichment analyses. As a whole, single TFs, TF cooperations, regulatory pathways and master regulators can further help to understand the underlying mechanisms that differ between the individual regulatory programs in different cell lines.



## Bibliography

- [1] Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenko VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, **459**(7243):108–112.
- [2] Navarro C, Lopez FJ, Cano C, Garcia-Alcalde F, Blanco A: **CisMiner: genome-wide in-silico cis-regulatory module prediction by fuzzy itemset mining.** *PLoS ONE* 2014, **9**(9):e108065.
- [3] Jankowski A, Prabhakar S, Tiuryn J: **TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers.** *BMC Genomics* 2014, **15**:208.
- [4] Deyneko IV, Kel AE, Kel-Margoulis OV, Deineko EV, Wingender E, Weiss S: **MatrixCatch—a novel tool for the recognition of composite regulatory elements in promoters.** *BMC Bioinformatics* 2013, **14**:241.
- [5] Nandi S, Blais A, Ioshikhes I: **Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors.** *Nucleic Acids Res.* 2013, **41**(19):8822–8841.
- [6] Ha N, Polychronidou M, Lohmann I: **COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets.** *PLoS ONE* 2012, **7**(12):e52055.
- [7] Sun H, Guns T, Fierro AC, Thorrez L, Nijssen S, Marchal K: **Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection.** *Nucleic Acids Res.* 2012, **40**(12):e90.
- [8] Hu Z, Hu B, Collins JF: **Prediction of synergistic transcription factors by function conservation.** *Genome Biol.* 2007, **8**(12):R257.
- [9] Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res.* 2003, **31**(13):3666–3668.
- [10] Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17**(10):878–889.

- [11] Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19 Suppl 1**:292–301.
- [12] Mysickova A, Vingron M: **Detection of interacting transcription factors in human tissues using predicted DNA binding affinity.** *BMC Genomics* 2012, **13 Suppl 1**:S2.
- [13] Girgis HZ, Ovcharenko I: **Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs.** *BMC Bioinformatics* 2012, **13**:25.
- [14] Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E: **TRANSCompel: a database on composite regulatory elements in eukaryotic genes.** *Nucleic Acids Res.* 2002, **30**:332–334.
- [15] Wong KC: **MotifHyades: expectation maximization for de novo DNA motif pair discovery on paired sequences.** *Bioinformatics* 2017, **33**(19):3028–3035.
- [16] Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R: **Molecular Biology of the Gene - International Edition** 2008.
- [17] Berg J, Guglielmi J, Tymoczko J, Held A, Stryer L, Kuhlmann-Krieg S, Pfeiffer-Guglielmi B, Seidler L, Vogel S, von der Saal K, et al.: **Biochemie** 2003, [<https://books.google.de/books?id=LPTdAAAACAAJ>].
- [18] Wingender E: **Gene Regulation in Eukaryotes** 1993, [<https://books.google.de/books?id=H3zwAAAAMAAJ>].
- [19] Mora A, Sandve GK, Gabrielsen OS, Eskeland R: **In the loop: promoter-enhancer interactions and bioinformatics.** *Brief. Bioinformatics* 2016, **17**(6):980–995.
- [20] Zhao C, Li X, Hu H: **PETModule: a motif module based approach for enhancer target gene prediction.** *Sci Rep* 2016, **6**:30043.
- [21] van Arensbergen J, van Steensel B, Bussemaker HJ: **In search of the determinants of enhancer-promoter interaction specificity.** *Trends Cell Biol.* 2014, **24**(11):695–702.
- [22] Ong CT, Corces VG: **Enhancer function: new insights into the regulation of tissue-specific gene expression.** *Nat. Rev. Genet.* 2011, **12**(4):283–293.
- [23] He B, Chen C, Teng L, Tan K: **Global view of enhancer-promoter interactome in human cells.** *Proc. Natl. Acad. Sci. U.S.A.* 2014, **111**(21):E2191–2199.

- [24] Matharu N, Ahituv N: **Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease.** *PLoS Genet.* 2015, **11**(12):e1005640.
- [25] Wingender E, Schoeps T, Donitz J: **TFClass: an expandable hierarchical classification of human transcription factors.** *Nucleic Acids Res.* 2013, **41**(Database issue):D165–170.
- [26] Wingender E: **Classification Scheme of Eukaryotic Transcription Factors.** *Molecular Biology* 1997, **31**(4):483–497.
- [27] Dang TKL, Meckbach C, Tacke R, Waack S, Gültas M: **A Novel Sequence-Based Feature for the Identification of DNA-Binding Sites in Proteins Using Jensen–Shannon Divergence.** *Entropy* 2016, **18**(10), [<http://www.mdpi.com/1099-4300/18/10/379>].
- [28] de Wit E, de Laat W: **A decade of 3C technologies: insights into nuclear organization.** *Genes Dev.* 2012, **26**:11–24.
- [29] Wingender E: **Compilation of transcription regulating proteins.** *Nucleic Acids Res.* 1988, **16**(5):1879–1902.
- [30] Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res.* 1996, **24**:238–241.
- [31] **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306**(5696):636–640.
- [32] Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database.** *Nucleic Acids Res.* 2003, **31**:51–54.
- [33] Breitkreutz BJ, Stark C, Tyers M: **The GRID: the General Repository for Interaction Datasets.** *Genome Biol.* 2003, **4**(3):R23.
- [34] Snel B, Lehmann G, Bork P, Huynen MA: **STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene.** *Nucleic Acids Res.* 2000, **28**(18):3442–3444.
- [35] Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res.* 2003, **31**(13):3576–3579.

- [36] Jiang M, Anderson J, Gillespie J, Mayne M: **uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts.** *BMC Bioinformatics* 2008, **9**:192.
- [37] Cover TM, Thomas JA: **Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)** 2006.
- [38] Timme N, Alford W, Flecker B, Beggs JM: **Synergy, redundancy, and multivariate information measures: an experimentalist's perspective.** *J Comput Neurosci* 2014, **36**(2):119–140.
- [39] Han TS: **Nonnegative Entropy Measures of Multivariate Symmetric Correlations.** *Information and Control* 1978, **36**:133–156.
- [40] Church KW, Hanks P: **Word Association Norms, Mutual Information, and Lexicography.** *Comput. Linguist.* 1990, **16**:22–29, [<http://dl.acm.org/citation.cfm?id=89086.89095>].
- [41] Bouma G: **Normalized (pointwise) mutual information in collocation extraction.** In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009, Volume Normalized*, Tübingen 2009:31–40.
- [42] Aji S, Kaimal MR: **Document summarization using positive pointwise mutual information.** *CoRR* 2012, **abs/1205.1638**, [<http://arxiv.org/abs/1205.1638>].
- [43] Meckbach C, Tacke R, Hua X, Waack S, Wingender E, Gultas M: **PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information.** *BMC Bioinformatics* 2015, **16**:400.
- [44] Meckbach C, Wingender E, Gultas M: **Removing Background Co-occurrences of Transcription Factor Binding Sites Greatly Improves the Prediction of Specific Transcription Factor Cooperations.** *Front Genet* 2018, **9**:189.
- [45] Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Res.* 2004, **32**(Database issue):D493–496.
- [46] Hannenhalli S, Levy S: **Predicting transcription factor synergism.** *Nucleic Acids Res.* 2002, **30**(19):4278–4284.
- [47] Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, Myers RM, Weng Z: **Functional analysis of transcription factor binding sites in human promoters.** *Genome Biol.* 2012, **13**(9):R50.

- [48] Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics* 2008, **24**(3):333–340.
- [49] Joshi H, Nord SH, Frigessi A, Børresen-Dale AL, Kristensen VN: **Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes.** *BMC Genomics* 2012, **13**:199.
- [50] Zeidler S, Meckbach C, Tacke R, Raad FS, Roa A, Uchida S, Zimmermann WH, Wingender E, Gultas M: **Computational Detection of Stage-Specific Transcription Factor Clusters during Heart Development.** *Front Genet* 2016, **7**:33.
- [51] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C: **STRING v10: protein-protein interaction networks, integrated over the tree of life.** *Nucleic Acids Res.* 2015, **43**(Database issue):D447–452.
- [52] Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M: **The BioGRID interaction database: 2015 update.** *Nucleic Acids Res.* 2015, **43**(Database issue):D470–478.
- [53] Kaczynski J, Cook T, Urrutia R: **Sp1- and Krüppel-like transcription factors.** *Genome Biol.* 2003, **4**(2):206.
- [54] Beishline K, Azizkhan-Clifford J: **Sp1 and the 'hallmarks of cancer'.** *FEBS J.* 2015, **282**(2):224–258.
- [55] Goenka S, Kaplan MH: **Transcriptional regulation by STAT6.** *Immunol. Res.* 2011, **50**:87–96.
- [56] Obika S, Reddy SY, Bruice TC: **Sequence specific DNA binding of Ets-1 transcription factor: molecular dynamics study on the Ets domain–DNA complexes.** *J. Mol. Biol.* 2003, **331**(2):345–359.
- [57] Hess J, Angel P, Schorpp-Kistner M: **AP-1 subunits: quarrel and harmony among siblings.** *J. Cell. Sci.* 2004, **117**(Pt 25):5965–5973.
- [58] Karin M, Liu Zg, Zandi E: **AP-1 function and regulation.** *Curr. Opin. Cell Biol.* 1997, **9**(2):240–246.
- [59] Block KL, Shou Y, Poncz M: **An Ets/Sp1 interaction in the 5'-flanking region of the megakaryocyte-specific alpha IIb gene appears to stabilize Sp1 binding and is essential for expression of this TATA-less gene.** *Blood* 1996, **88**(6):2071–2080.

- [60] Sahoo A, Lee CG, Jash A, Son JS, Kim G, Kwon HK, So JS, Im SH: **Stat6 and c-Jun mediate Th2 cell-specific IL-24 gene expression.** *J. Immunol.* 2011, **186**(7):4098–4109.
- [61] Switzer CH, Cheng RY, Ridnour LA, Glynn SA, Ambs S, Wink DA: **Ets-1 is a transcriptional mediator of oncogenic nitric oxide signaling in estrogen receptor-negative breast cancer.** *Breast Cancer Res.* 2012, **14**(5):R125.
- [62] Alvira CM: **Nuclear factor-kappa-B signaling in lung development and disease: one pathway, numerous functions.** *Birth Defects Res. Part A Clin. Mol. Teratol.* 2014, **100**(3):202–216.
- [63] Zhang W, Grivennikov SI: **Top Notch cancer stem cells by paracrine NF- $\kappa$ B signaling in breast cancer.** *Breast Cancer Res.* 2013, **15**(5):316.
- [64] Takai N, Miyazaki T, Nishida M, Shang S, Nasu K, Miyakawa I: **Clinical relevance of Elf-1 overexpression in endometrial carcinoma.** *Gynecol. Oncol.* 2003, **89**(3):408–413.
- [65] Ecevit O, Khan MA, Goss DJ: **Kinetic analysis of the interaction of b/HLH/Z transcription factors Myc, Max, and Mad with cognate DNA.** *Biochemistry* 2010, **49**(12):2627–2635.
- [66] Xu J, Chen Y, Olopade OI: **MYC and Breast Cancer.** *Genes Cancer* 2010, **1**(6):629–640.
- [67] Chen Y, Xu J, Borowicz S, Collins C, Huo D, Olopade OI: **c-Myc activates BRCA1 gene expression through distal promoter elements in breast cancer cells.** *BMC Cancer* 2011, **11**:246.
- [68] Bindra RS, Gibson SL, Meng A, Westermarck U, Jasin M, Pierce AJ, Bristow RG, Classon MK, Glazer PM: **Hypoxia-induced down-regulation of BRCA1 expression by E2Fs.** *Cancer Res.* 2005, **65**(24):11597–11604.
- [69] Champion CG, Labrie M, Grosset AA, St-Pierre Y: **The CCAAT/enhancer-binding protein beta-2 isoform (CEBP $\beta$ -2) upregulates galectin-7 expression in human breast cancer cells.** *PLoS ONE* 2014, **9**(5):e95087.
- [70] Shah SN, Cope L, Poh W, Belton A, Roy S, Talbot CC, Sukumar S, Huso DL, Resar LM: **HMGA1: a master regulator of tumor progression in triple-negative breast cancer cells.** *PLoS ONE* 2013, **8**(5):e63419.
- [71] George OL, Ness SA: **Situational awareness: regulation of the myb transcription factor in differentiation, the cell cycle and oncogenesis.** *Cancers (Basel)* 2014, **6**(4):2049–2071.



- [72] Conway JR, Lex A, Gehlenborg N: **UpSetR: an R package for the visualization of intersecting sets and their properties.** *Bioinformatics* 2017, **33**(18):2938–2940.
- [73] Whalen S, Truty RM, Pollard KS: **Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin.** *Nat. Genet.* 2016, **48**(5):488–496.
- [74] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell* 2014, **159**(7):1665–1680.
- [75] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, Osborne CS: **Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C.** *Nat. Genet.* 2015, **47**(6):598–606.
- [76] Rasheed ZA, Saleem A, Ravee Y, Pandolfi PP, Rubin EH: **The topoisomerase I-binding RING protein, topors, is associated with promyelocytic leukemia nuclear bodies.** *Exp. Cell Res.* 2002, **277**(2):152–160.
- [77] Bredel M, Bredel C, Juric D, Harsh GR, Vogel H, Recht LD, Sikic BI: **High-resolution genome-wide mapping of genetic alterations in human glial brain tumors.** *Cancer Res.* 2005, **65**(10):4088–4096.
- [78] Haluska P, Saleem A, Rasheed Z, Ahmed F, Su EW, Liu LF, Rubin EH: **Interaction between human topoisomerase I and a novel RING finger/arginine-serine protein.** *Nucleic Acids Res.* 1999, **27**(12):2538–2544.
- [79] Saleem A, Dutta J, Malegaonkar D, Rasheed F, Rasheed Z, Rajendra R, Marshall H, Luo M, Li H, Rubin EH: **The topoisomerase I- and p53-binding protein topors is differentially expressed in normal and malignant human tissues and may function as a tumor suppressor.** *Oncogene* 2004, **23**(31):5293–5300.
- [80] Weger S, Hammer E, Heilbronn R: **Topors acts as a SUMO-1 E3 ligase for p53 in vitro and in vivo.** *FEBS Lett.* 2005, **579**(22):5007–5012.
- [81] Gonzalez P, Garcia-Castro M, Reguero JR, Batalla A, Ordonez AG, Palop RL, Lozano I, Montes M, Alvarez V, Coto E: **The Pro279Leu variant in the transcription factor MEF2A is associated with myocardial infarction.** *J. Med. Genet.* 2006, **43**(2):167–169.

- [82] Bai X, Wu L, Liang T, Liu Z, Li J, Li D, Xie H, Yin S, Yu J, Lin Q, Zheng S: **Overexpression of myocyte enhancer factor 2 and histone hyperacetylation in hepatocellular carcinoma.** *J. Cancer Res. Clin. Oncol.* 2008, **134**:83–91.
- [83] Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, Onofrio R, Pugh TJ, Rajkumar SV, Ramos AH, Siegel DS, Sivachenko A, Stewart AK, Trudel S, Vij R, Voet D, Winckler W, Zimmerman T, Carpten J, Trent J, Hahn WC, Garraway LA, Meyerson M, Lander ES, Getz G, Golub TR: **Initial genome sequencing and analysis of multiple myeloma.** *Nature* 2011, **471**(7339):467–472.
- [84] Wu Y, Dey R, Han A, Jayathilaka N, Philips M, Ye J, Chen L: **Structure of the MADS-box/MEF2 domain of MEF2A bound to DNA and its implication for myocardin recruitment.** *J. Mol. Biol.* 2010, **397**(2):520–533.
- [85] Gonzalez P, Alvarez V, Menendez M, Lahoz CH, Martinez C, Corao AI, Calatayud MT, Pena J, Garcia-Castro M, Coto E: **Myocyte enhancing factor-2A in Alzheimer's disease: genetic analysis and association with MEF2A-polymorphisms.** *Neurosci. Lett.* 2007, **411**:47–51.
- [86] Thai MV, Guruswamy S, Cao KT, Pessin JE, Olson AL: **Myocyte enhancer factor 2 (MEF2)-binding site is required for GLUT4 gene expression in transgenic mice. Regulation of MEF2 DNA binding activity in insulin-deficient diabetes.** *J. Biol. Chem.* 1998, **273**(23):14285–14292.
- [87] Elhawari S, Al-Boudari O, Muiya P, Khalak H, Andres E, Al-Shahid M, Al-Dosari M, Meyer BF, Al-Mohanna F, Dzimir N: **A study of the role of the Myocyte-specific Enhancer Factor-2A gene in coronary artery disease.** *Atherosclerosis* 2010, **209**:152–154.
- [88] Pon JR, Marra MA: **MEF2 transcription factors: developmental regulators and emerging cancer genes.** *Oncotarget* 2016, **7**(3):2297–2312.
- [89] Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.** *Genome Res.* 2012, **22**(9):1798–1812.
- [90] Lozzio CB, Lozzio BB: **Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome.** *Blood* 1975, **45**(3):321–334.

- [91] Guo B, Odgren PR, van Wijnen AJ, Last TJ, Nickerson J, Penman S, Lian JB, Stein JL, Stein GS: **The nuclear matrix protein NMP-1 is the transcription factor YY1.** *Proc. Natl. Acad. Sci. U.S.A.* 1995, **92**(23):10526–10530.
- [92] Moriuchi M, Moriuchi H: **YY1 transcription factor down-regulates expression of CCR5, a major coreceptor for HIV-1.** *J. Biol. Chem.* 2003, **278**(15):13003–13007.
- [93] Lichy JH, Majidi M, Elbaum J, Tsai MM: **Differential expression of the human ST5 gene in HeLa-fibroblast hybrid cell lines mediated by YY1: evidence that YY1 plays a part in tumor suppression.** *Nucleic Acids Res.* 1996, **24**(23):4700–4708.
- [94] Gronroos E, Terentiev AA, Punga T, Ericsson J: **YY1 inhibits the activation of the p53 tumor suppressor in response to genotoxic stress.** *Proc. Natl. Acad. Sci. U.S.A.* 2004, **101**(33):12165–12170.
- [95] Atchison M, Basu A, Zaprazna K, Papanasi M: **Mechanisms of Yin Yang 1 in oncogenesis: the importance of indirect effects.** *Crit Rev Oncog* 2011, **16**(3-4):143–161.
- [96] Erkeland SJ, Valkhof M, Heijmans-Antonissen C, Delwel R, Valk PJ, Hermans MH, Touw IP: **The gene encoding the transcriptional regulator Yin Yang 1 (YY1) is a myeloid transforming gene interfering with neutrophilic differentiation.** *Blood* 2003, **101**(3):1111–1117.
- [97] Grubach L, Juhl-Christensen C, Rethmeier A, Olesen LH, Aggerholm A, Hokland P, Ostergaard M: **Gene expression profiling of Polycomb, Hox and Meis genes in patients with acute myeloid leukaemia.** *Eur. J. Haematol.* 2008, **81**(2):112–122.
- [98] Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, Guo YE, Hnisz D, Jaenisch R, Bradner JE, Gray NS, Young RA: **YY1 Is a Structural Regulator of Enhancer-Promoter Loops.** *Cell* 2017, **171**(7):1573–1588.
- [99] Wang J, Wu X, Wei C, Huang X, Ma Q, Huang X, Faiola F, Guallar D, Fidalgo M, Huang T, Peng D, Chen L, Yu H, Li X, Sun J, Liu X, Cai X, Chen X, Wang L, Ren J, Wang J, Ding J: **YY1 Positively Regulates Transcription by Targeting Promoters and Super-Enhancers through the BAF Complex in Embryonic Stem Cells.** *Stem Cell Reports* 2018, **10**(4):1324–1339.
- [100] Kravets A, Hu Z, Miralem T, Torno MD, Maines MD: **Biliverdin reductase, a novel regulator for induction of activating transcription factor-2 and heme oxygenase-1.** *J. Biol. Chem.* 2004, **279**(19):19916–19923.

- [101] Ouwens DM, de Ruiter ND, van der Zon GC, Carter AP, Schouten J, van der Burgt C, Kooistra K, Bos JL, Maassen JA, van Dam H: **Growth factors can activate ATF2 via a two-step mechanism: phosphorylation of Thr71 through the Ras-MEK-ERK pathway and of Thr69 through RalGDS-Src-p38.** *EMBO J.* 2002, **21**(14):3782–3793.
- [102] Berger AJ, Kluger HM, Li N, Kielhorn E, Halaban R, Ronai Z, Rimm DL: **Sub-cellular localization of activating transcription factor 2 in melanoma specimens predicts patient survival.** *Cancer Res.* 2003, **63**(23):8103–8107.
- [103] Makino C, Sano Y, Shinagawa T, Millar JB, Ishii S: **Sin1 binds to both ATF-2 and p38 and enhances ATF-2-dependent transcription in an SAPK signaling pathway.** *Genes Cells* 2006, **11**(11):1239–1251.
- [104] Woo IS, Kohno T, Inoue K, Ishii S, Yokota J: **Infrequent mutations of the activating transcription factor-2 gene in human lung cancer, neuroblastoma and breast cancer.** *Int. J. Oncol.* 2002, **20**(3):527–531.
- [105] Tsai EY, Jain J, Pesavento PA, Rao A, Goldfeld AE: **Tumor necrosis factor alpha gene regulation in activated T cells involves ATF-2/Jun and NFATp.** *Mol. Cell Biol.* 1996, **16**(2):459–467.
- [106] Chen KC, Chiou YL, Chang LS: **JNK1/c-Jun and p38 alpha MAPK/ATF-2 pathways are responsible for upregulation of Fas/FasL in human chronic myeloid leukemia K562 cells upon exposure to Taiwan cobra phospholipase A2.** *J. Cell Biochem.* 2009, **108**(3):612–620.
- [107] Panne D, Maniatis T, Harrison SC: **Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon- $\beta$  enhancer.** *The EMBO Journal* 2004, **23**(22):4384–4393, [<http://emboj.embopress.org/content/23/22/4384>].
- [108] Mu X, Springer JE, Bowser R: **FAC1 expression and localization in motor neurons of developing, adult, and amyotrophic lateral sclerosis spinal cord.** *Exp. Neurol.* 1997, **146**:17–24.
- [109] Jordan-Sciutto KL, Dragich JM, Caltagarone J, Hall DJ, Bowser R: **Fetal Alz-50 clone 1 (FAC1) protein interacts with the Myc-associated zinc finger protein (ZF87/MAZ) and alters its transcriptional activity.** *Biochemistry* 2000, **39**(12):3206–3215.
- [110] Buganim Y, Goldstein I, Lipson D, Milyavsky M, Polak-Charcon S, Mardoukh C, Solomon H, Kalo E, Madar S, Brosh R, Perelman M, Navon R, Goldfinger N, Barshack I, Yakhini Z, Rotter V: **A novel translocation breakpoint within the BPTF gene is associated with a pre-malignant phenotype.** *PLoS ONE* 2010, **5**(3):e9657.

- [111] Grinberg-Rashi H, Ofek E, Perelman M, Skarda J, Yaron P, Hajduch M, Jacob-Hirsch J, Amariglio N, Krupsky M, Simansky DA, Ram Z, Pfeffer R, Galernter I, Steinberg DM, Ben-Dov I, Rechavi G, Izraeli S: **The expression of three genes in primary non-small cell lung cancer is associated with metastatic spread to the brain.** *Clin. Cancer Res.* 2009, **15**(5):1755–1761.
- [112] Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD, Patel DJ: **Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF.** *Nature* 2006, **442**(7098):91–95.
- [113] Bowser R, Reilly S: **Expression of FAC1 in activated microglia during Alzheimer's disease.** *Neurosci. Lett.* 1998, **253**(3):163–166.
- [114] Barak O, Lazzaro MA, Lane WS, Speicher DW, Picketts DJ, Shiekhattar R: **Isolation of human NURF: a regulator of Engrailed gene expression.** *EMBO J.* 2003, **22**(22):6089–6100.
- [115] Xu B, Cai L, Butler JM, Chen D, Lu X, Allison DF, Lu R, Rafii S, Parker JS, Zheng D, Wang GG: **The Chromatin Remodeler BPTF Activates a Stemness Gene-Expression Program Essential for the Maintenance of Adult Hematopoietic Stem Cells.** *Stem Cell Reports* 2018, **10**(3):675–683.
- [116] Bjørkhaug L, Ye H, Horikawa Y, Søvik O, Molven A, Njølstad PR: **MODY associated with two novel hepatocyte nuclear factor-1alpha loss-of-function mutations (P112L and Q466X).** *Biochem. Biophys. Res. Commun.* 2000, **279**(3):792–798.
- [117] Barrio R, Bellanne-Chantelot C, Moreno JC, Morel V, Calle H, Alonso M, Mustieles C: **Nine novel mutations in maturity-onset diabetes of the young (MODY) candidate genes in 22 Spanish families.** *J. Clin. Endocrinol. Metab.* 2002, **87**(6):2532–2539.
- [118] Bonham K, Ritchie SA, Dehm SM, Snyder K, Boyd FM: **An alternative, human SRC promoter and its regulation by hepatic nuclear factor-1alpha.** *J. Biol. Chem.* 2000, **275**(48):37604–37611.
- [119] Rebouissou S, Imbeaud S, Balabaud C, Boulanger V, Bertrand-Michel J, Terce F, Auffray C, Bioulac-Sage P, Zucman-Rossi J: **HNF1alpha inactivation promotes lipogenesis in human hepatocellular adenoma independently of SREBP-1 and carbohydrate-response element-binding protein (ChREBP) activation.** *J. Biol. Chem.* 2007, **282**(19):14437–14446.
- [120] Chiu KC, Chuang LM, Ryu JM, Tsai GP, Saad MF: **The I27L amino acid polymorphism of hepatic nuclear factor-1alpha is associated with insulin resistance.** *J. Clin. Endocrinol. Metab.* 2000, **85**(6):2178–2183.

- [121] Vaxillaire M, Abderrahmani A, Boutin P, Bailleul B, Froguel P, Yaniv M, Pontoglio M: **Anatomy of a homeoprotein revealed by the analysis of human MODY3 mutations.** *J. Biol. Chem.* 1999, **274**(50):35639–35646.
- [122] Hagiwara N, Mechanic LE, Trivers GE, Cawley HL, Taga M, Bowman ED, Kummamoto K, He P, Bernard M, Doja S, Miyashita M, Tajiri T, Sasajima K, Nomura T, Makino H, Takahashi K, Hussain SP, Harris CC: **Quantitative detection of p53 mutations in plasma DNA from tobacco smokers.** *Cancer Res.* 2006, **66**(16):8309–8317.
- [123] Matoba S, Kang JG, Patino WD, Wragg A, Boehm M, Gavrilova O, Hurley PJ, Bunz F, Hwang PM: **p53 regulates mitochondrial respiration.** *Science* 2006, **312**(5780):1650–1653.
- [124] Okoshi R, Ozaki T, Yamamoto H, Ando K, Koida N, Ono S, Koda T, Kamijo T, Nakagawara A, Kizaki H: **Activation of AMP-activated protein kinase induces p53-dependent apoptotic cell death in response to energetic stress.** *J. Biol. Chem.* 2008, **283**(7):3979–3987.
- [125] Jung P, Verdoodt B, Bailey A, Yates JR, Menssen A, Hermeking H: **Induction of cullin 7 by DNA damage attenuates p53 function.** *Proc. Natl. Acad. Sci. U.S.A.* 2007, **104**(27):11388–11393.
- [126] Liu Y, Bodmer WF: **Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines.** *Proc. Natl. Acad. Sci. U.S.A.* 2006, **103**(4):976–981.
- [127] Hollstein M, Sidransky D, Vogelstein B, Harris CC: **p53 mutations in human cancers.** *Science* 1991, **253**(5015):49–53.
- [128] Golubovskaya VM, Finch R, Zheng M, Kurenova EV, Cance WG: **The 7-amino-acid site in the proline-rich region of the N-terminal domain of p53 is involved in the interaction with FAK and is critical for p53 functioning.** *Biochem. J.* 2008, **411**:151–160.
- [129] Huang R, Liao X, Li Q: **Identification of key pathways and genes in TP53 mutation acute myeloid leukemia: evidence from bioinformatics analysis.** *Onco Targets Ther* 2018, **11**:163–173.
- [130] Kadia TM, Jain P, Ravandi F, Garcia-Manero G, Andreef M, Takahashi K, Borthakur G, Jabbour E, Konopleva M, Daver NG, Dinardo C, Pierce S, Kanagal-Shamanna R, Patel K, Estrov Z, Cortes J, Kantarjian HM: **TP53 mutations in newly diagnosed acute myeloid leukemia: Clinicomolecular characteristics, response to therapy, and outcomes.** *Cancer* 2016, **122**(22):3484–3491.

- [131] Chiaretti S, Brugnoletti F, Tavolaro S, Bonina S, Paoloni F, Marinelli M, Patten N, Bonifacio M, Kropp MG, Sica S, Guarini A, Foa R: **TP53 mutations are frequent in adult acute lymphoblastic leukemia cases negative for recurrent fusion genes and correlate with poor response to induction therapy.** *Haematologica* 2013, **98**(5):59–61.
- [132] Usuda J, Inomata M, Fukumoto H, Iwamoto Y, Suzuki T, Kuh HJ, Fukuoka K, Kato H, Saijo N, Nishio K: **Restoration of p53 gene function in 12-O-tetradecanoylphorbol 13-acetate-resistant human leukemia K562/TPA cells.** *Int. J. Oncol.* 2003, **22**:81–86.
- [133] Law JC, Ritke MK, Yalowich JC, Leder GH, Ferrell RE: **Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line.** *Leuk. Res.* 1993, **17**(12):1045–1050.
- [134] Magdinier F, Dalla Venezia N, Lenoir GM, Frappart L, Dante R: **BRCA1 expression during prenatal development of the human mammary gland.** *Oncogene* 1999, **18**(27):4039–4043.
- [135] Xian Ma Y, Fan S, Xiong J, Yuan RQ, Meng Q, Gao M, Goldberg ID, Fuqua SA, Pestell RG, Rosen EM: **Role of BRCA1 in heat shock response.** *Oncogene* 2003, **22**:10–27.
- [136] Wallin J, Eibel H, Neubuser A, Wilting J, Koseki H, Balling R: **Pax1 is expressed during development of the thymus epithelium and is required for normal T-cell maturation.** *Development* 1996, **122**:23–30.
- [137] Wheat W, Fitzsimmons D, Lennox H, Krautkramer SR, Gentile LN, McIntosh LP, Hagman J: **The highly conserved beta-hairpin of the paired DNA-binding domain is required for assembly of Pax-Ets ternary complexes.** *Mol. Cell. Biol.* 1999, **19**(3):2231–2241.
- [138] Salomon R, Tellier AL, Attie-Bitach T, Amiel J, Vekemans M, Lyonnet S, Dureau P, Niaudet P, Gubler MC, Broyer M: **PAX2 mutations in oligomeganephronia.** *Kidney Int.* 2001, **59**(2):457–462.
- [139] Hueber PA, Waters P, Clark P, Clarke P, Eccles M, Goodyer P: **PAX2 inactivation enhances cisplatin-induced apoptosis in renal carcinoma cells.** *Kidney Int.* 2006, **69**(7):1139–1145.
- [140] Luu VD, Boysen G, Struckmann K, Casagrande S, von Teichman A, Wild PJ, Sulser T, Schraml P, Moch H: **Loss of VHL and hypoxia provokes PAX2 up-regulation in clear cell renal cell carcinoma.** *Clin. Cancer Res.* 2009, **15**(10):3297–3304.

- [141] Thomas R, Sanna-Cherchi S, Warady BA, Furth SL, Kaskel FJ, Gharavi AG: **HNF1B and PAX2 mutations are a common cause of renal hypodysplasia in the CKiD cohort.** *Pediatr. Nephrol.* 2011, **26**(6):897–903.
- [142] Tung CS, Mok SC, Tsang YT, Zu Z, Song H, Liu J, Deavers MT, Malpica A, Wolf JK, Lu KH, Gershenson DM, Wong KK: **PAX2 expression in low malignant potential ovarian tumors and low-grade ovarian serous carcinomas.** *Mod. Pathol.* 2009, **22**(9):1243–1250.
- [143] Walton KL, Johnson KE, Harrison CA: **Targeting TGF- $\beta$  Mediated SMAD Signaling for the Prevention of Fibrosis.** *Front Pharmacol* 2017, **8**:461.
- [144] Foster KW, Frost AR, McKie-Bell P, Lin CY, Engler JA, Grizzle WE, Ruppert JM: **Increase of GSK3 $\beta$  messenger RNA and protein expression during progression of breast cancer.** *Cancer Res.* 2000, **60**(22):6488–6495.
- [145] McCormick SM, Eskin SG, McIntire LV, Teng CL, Lu CM, Russell CG, Chittur KK: **DNA microarray reveals changes in gene expression of shear stressed human umbilical vein endothelial cells.** *Proc. Natl. Acad. Sci. U.S.A.* 2001, **98**(16):8955–8960.
- [146] Yasunaga J, Taniguchi Y, Nosaka K, Yoshida M, Satou Y, Sakai T, Mitsuya H, Matsuoka M: **Identification of aberrantly methylated genes in association with adult T-cell leukemia.** *Cancer Res.* 2004, **64**(17):6002–6009.
- [147] Wei D, Gong W, Kanai M, Schlunk C, Wang L, Yao JC, Wu TT, Huang S, Xie K: **Drastic down-regulation of Krüppel-like factor 4 expression is critical in human gastric cancer development and progression.** *Cancer Res.* 2005, **65**(7):2746–2754.
- [148] Jenkins TD, Opitz OG, Okano J, Rustgi AK: **Transactivation of the human keratin 4 and Epstein-Barr virus ED-L2 promoters by gut-enriched Krüppel-like factor.** *J. Biol. Chem.* 1998, **273**(17):10747–10754.
- [149] Luo A, Kong J, Hu G, Liew CC, Xiong M, Wang X, Ji J, Wang T, Zhi H, Wu M, Liu Z: **Discovery of Ca<sup>2+</sup>-relevant and differentiation-associated genes downregulated in esophageal squamous cell carcinoma using cDNA microarray.** *Oncogene* 2004, **23**(6):1291–1299.
- [150] Zhao W, Hisamuddin IM, Nandan MO, Babbin BA, Lamb NE, Yang VW: **Identification of Krüppel-like factor 4 as a potential tumor suppressor gene in colorectal cancer.** *Oncogene* 2004, **23**(2):395–402.
- [151] Lin L, Han Q, Xiong Y, Li T, Liu Z, Xu H, Wu Y, Wang N, Liu X: **Krüppel-like-factor 4 Attenuates Lung Fibrosis via Inhibiting Epithelial-mesenchymal Transition.** *Sci Rep* 2017, **7**:15847.



- [152] Scavuzzo MA, Chmielowiec J, Yang D, Wamble K, Chaboub LS, Duraine L, Tepe B, Glasgow SM, Arenkiel BR, Brou C, Deneen B, Borowiak M: **Pancreatic Cell Fate Determination Relies on Notch Ligand Trafficking by NFIA.** *Cell Rep* 2018, **25**(13):3811–3827.
- [153] Dizier MH, Margaritte-Jeannin P, Madore AM, Moffatt M, Brossard M, Lavielle N, Sarnowski C, Just J, Cookson W, Lathrop M, Laprise C, Bouzigon E, Demenais F: **The nuclear factor I/A (NFIA) gene is associated with the asthma plus rhinitis phenotype.** *J. Allergy Clin. Immunol.* 2014, **134**(3):576–582.
- [154] Wu Y, Zhang J, Hou S, Cheng Z, Yuan M: **Non-small cell lung cancer: miR-30d suppresses tumor invasion and migration by directly targeting NFIB.** *Biotechnol. Lett.* 2017, **39**(12):1827–1834.
- [155] Zhu HX, Shi L, Zhang Y, Zhu YC, Bai CX, Wang XD, Zhou JB: **Myocyte enhancer factor 2D provides a cross-talk between chronic inflammation and lung cancer.** *J Transl Med* 2017, **15**:65.
- [156] Rahman I, MacNee W: **Oxidative stress and regulation of glutathione in lung inflammation.** *Eur. Respir. J.* 2000, **16**(3):534–554.
- [157] Xu X, Kwon OK, Shin IS, Mali JR, Harmalkar DS, Lim Y, Lee G, Lu Q, Oh SR, Ahn KS, Jeong HG, Lee K: **Novel benzofuran derivative DK-1014 attenuates lung inflammation via blocking of MAPK/AP-1 and AKT/mTOR signaling in vitro and in vivo.** *Sci Rep* 2019, **9**:862.
- [158] Bergsagel PL, Kuehl WM: **Chromosome translocations in multiple myeloma.** *Oncogene* 2001, **20**(40):5611–5622.
- [159] Morito N, Yoh K, Fujioka Y, Nakano T, Shimohata H, Hashimoto Y, Yamada A, Maeda A, Matsuno F, Hata H, Suzuki A, Imagawa S, Mitsuya H, Esumi H, Koyama A, Yamamoto M, Mori N, Takahashi S: **Overexpression of c-Maf contributes to T-cell lymphoma in both mice and human.** *Cancer Res.* 2006, **66**(2):812–819.
- [160] Ho IC, Hodge MR, Rooney JW, Glimcher LH: **The proto-oncogene c-maf is responsible for tissue-specific expression of interleukin-4.** *Cell* 1996, **85**(7):973–983.
- [161] Jamieson RV, Perveen R, Kerr B, Carette M, Yardley J, Heon E, Wirth MG, van Heyningen V, Donnai D, Munier F, Black GC: **Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma.** *Hum. Mol. Genet.* 2002, **11**:33–42.

- [162] Cao S, Liu J, Chesi M, Bergsagel PL, Ho IC, Donnelly RP, Ma X: **Differential regulation of IL-12 and IL-10 gene expression in macrophages by the basic leucine zipper transcription factor c-Maf fibrosarcoma.** *J. Immunol.* 2002, **169**(10):5715–5725.
- [163] Shaw JG, Vaughan A, Dent AG, O'Hare PE, Goh F, Bowman RV, Fong KM, Yang IA: **Biomarkers of progression of chronic obstructive pulmonary disease (COPD).** *J Thorac Dis* 2014, **6**(11):1532–1547.
- [164] Bakre A, Wu W, Hiscox J, Spann K, Teng MN, Tripp RA: **Human respiratory syncytial virus non-structural protein NS1 modifies miR-24 expression via transforming growth factor- $\beta$ .** *J. Gen. Virol.* 2015, **96**(11):3179–3191.
- [165] Leng D, Huan C, Xie T, Liang J, Wang J, Dai H, Wang C, Jiang D: **Meta-analysis of genetic programs between idiopathic pulmonary fibrosis and sarcoidosis.** *PLoS ONE* 2013, **8**(8):e71059.
- [166] Shiseki M, Nagashima M, Pedoux RM, Kitahama-Shiseki M, Miura K, Okamura S, Onogi H, Higashimoto Y, Appella E, Yokota J, Harris CC: **p29ING4 and p28ING5 bind to p53 and p300, and enhance p53 activity.** *Cancer Res.* 2003, **63**(10):2373–2378.
- [167] Garkavtsev I, Kozin SV, Chernova O, Xu L, Winkler F, Brown E, Barnett GH, Jain RK: **The candidate tumour suppressor protein ING4 regulates brain tumour growth and angiogenesis.** *Nature* 2004, **428**(6980):328–332.
- [168] Nirodi C, Hart J, Dhawan P, Moon NS, Nepveu A, Richmond A: **The role of CDP in the negative regulation of CXCL1 gene expression.** *J. Biol. Chem.* 2001, **276**(28):26122–26131.
- [169] Auferio B, Neufeld EJ, Orkin SH: **Sequence-specific DNA binding of individual cut repeats of the human CCAAT displacement/cut homeodomain protein.** *Proc. Natl. Acad. Sci. U.S.A.* 1994, **91**(16):7757–7761.
- [170] Tosi S, Scherer SW, Giudici G, Czepulkowski B, Biondi A, Kearney L: **Delineation of multiple deleted regions in 7q in myeloid disorders.** *Genes Chromosomes Cancer* 1999, **25**(4):384–392.
- [171] O'Connor MJ, Stunkel W, Koh CH, Zimmermann H, Bernard HU: **The differentiation-specific factor CDP/Cut represses transcription and replication of human papillomaviruses through a conserved silencing element.** *J. Virol.* 2000, **74**:401–410.

- [172] Li S, Moy L, Pittman N, Shue G, Auferio B, Neufeld EJ, LeLeiko NS, Walsh MJ: **Transcriptional repression of the cystic fibrosis transmembrane conductance regulator gene, mediated by CCAAT displacement protein/cut homolog, is associated with histone deacetylation.** *J. Biol. Chem.* 1999, **274**(12):7803–7815.
- [173] Lievens PM, Donady JJ, Tufarelli C, Neufeld EJ: **Repressor activity of CCAAT displacement protein in HL-60 myeloid leukemia cells.** *J. Biol. Chem.* 1995, **270**(21):12745–12750.
- [174] Sun H, De Bie T, Storms V, Fu Q, Dhollander T, Lemmens K, Verstuyf A, De Moor B, Marchal K: **ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules.** *BMC Bioinformatics* 2009, **10 Suppl 1**:S30.
- [175] Hu Z, Gallo SM: **Identification of interacting transcription factors regulating tissue gene expression in human.** *BMC Genomics* 2010, **11**:49.
- [176] Klepper K, Sandve GK, Abul O, Johansen J, Drablos F: **Assessment of composite motif discovery methods.** *BMC Bioinformatics* 2008, **9**:123.
- [177] Yu X, Lin J, Zack DJ, Qian J: **Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues.** *Nucleic Acids Res.* 2006, **34**(17):4925–4936.
- [178] Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA: **A compilation of composite regulatory elements affecting gene transcription in vertebrates.** *Nucleic Acids Res.* 1995, **23**(20):4097–4103.

## **A. Appendix**

### **A.1. PC-TraFF: identification of potentially collaborating TFs using pointwise mutual information**

METHODOLOGY ARTICLE

Open Access



# PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information

Cornelia Meckbach<sup>1\*</sup>, Rebecca Tacke<sup>1</sup>, Xu Hua<sup>1</sup>, Stephan Waack<sup>2</sup>, Edgar Wingender<sup>1</sup> and Mehmet Gültas<sup>1\*</sup>

## Abstract

**Background:** Transcription factors (TFs) are important regulatory proteins that govern transcriptional regulation. Today, it is known that in higher organisms different TFs have to cooperate rather than acting individually in order to control complex genetic programs. The identification of these interactions is an important challenge for understanding the molecular mechanisms of regulating biological processes. In this study, we present a new method based on pointwise mutual information, PC-TraFF, which considers the genome as a document, the sequences as sentences, and TF binding sites (TFBSs) as words to identify interacting TFs in a set of sequences.

**Results:** To demonstrate the effectiveness of PC-TraFF, we performed a genome-wide analysis and a breast cancer-associated sequence set analysis for protein coding and miRNA genes. Our results show that in any of these sequence sets, PC-TraFF is able to identify important interacting TF pairs, for most of which we found support by previously published experimental results. Further, we made a pairwise comparison between PC-TraFF and three conventional methods. The outcome of this comparison study strongly suggests that all these methods focus on different important aspects of interaction between TFs and thus the pairwise overlap between any of them is only marginal.

**Conclusions:** In this study, adopting the idea from the field of linguistics in the field of bioinformatics, we develop a new information theoretic method, PC-TraFF, for the identification of potentially collaborating transcription factors based on the idiosyncrasy of their binding site distributions on the genome. The results of our study show that PC-TraFF can successfully identify known interacting TF pairs and thus its currently biologically unconfirmed predictions could provide new hypotheses for further experimental validation. Additionally, the comparison of the results of PC-TraFF with the results of previous methods demonstrates that different methods with their specific scopes can perfectly supplement each other. Overall, our analyses indicate that PC-TraFF is a time-efficient method where its algorithm has a tractable computational time and memory consumption.

The PC-TraFF server is freely accessible at <http://pctraff.bioinf.med.uni-goettingen.de/>

## Background

Transcription factors (TFs) are a special class of gene regulatory proteins binding to short DNA motifs, known as transcription factor binding sites (TFBS). These TFBSs are located in promoters, which are found around the transcription start site (TSS). The binding of TFs frequently occurs in a cooperative manner due to their functional collaboration which leads to cis-regulatory modules

(CRMs). These modules are important for an effective regulation of the transcriptional machinery, even if they are not enriched in the corresponding promoter regions. The collaboration of TFs might stem from synergistic or antagonistic interactions between homotypic as well as heterotypic TFs. Such collaborations are likely to have effect on gene specificity and flexibility of the controlling of gene transcription during, for instance, tissue development and differentiation [1–3]. Thus, identification of collaborating TFs is as crucial as the determination of enriched TFs in genomic sequences for understanding the molecular mechanisms of cellular regulation [1].

\*Correspondence: [cornelia.meckbach@bioinf.med.uni-goettingen.de](mailto:cornelia.meckbach@bioinf.med.uni-goettingen.de);  
[mehmet.gultas@bioinf.med.uni-goettingen.de](mailto:mehmet.gultas@bioinf.med.uni-goettingen.de)

<sup>1</sup>Institute of Bioinformatics, University of Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany

Full list of author information is available at the end of the article



Until now, several groups have published different studies for the identification of cis-regulatory modules, and based on those studies, a variety of computational algorithms have been developed to determine potential interactions between TFs according to their binding sites [4–15]. However, many of these studies require negative and/or positive control sets and demand prior knowledge about TF pairs [3, 5, 8, 11]. Further, most of these studies often use simple organisms or restricted genes or focus only on statistically overrepresented TFBSs in DNA sequences. As a result, they usually have limited success, and thus only detect a small number of all interacting TFs (see the review [16] for the success rates of different CRM-methods).

Large efforts have been made in the last few years to overcome the limited success of existing methods. In these cases, different methods have been utilized such as searching the DNA for clusters of binding sites, comparing function conservation between related species, and applying association rules as well as statistical methods like the hypergeometric or the permutation test [4, 7, 8, 17]. Navarro et al. [4] have presented the Fuzzy Clustering approach, which has been already applied by Pickert et al. [18], in association with the Top-Down Fuzzy Frequent-Pattern Tree algorithm to detect significantly co-occurring TFBSs based on their locations on the DNA. Na et al. [8], have published in their study a co-occurring pattern search (COPS) combining association rules with a Markov model and only focusing on a predefined TF in simple organisms. However the scope of applicability of both methods is strongly limited due to their very high running time and memory consumption. As an example, the examination of the human genome is problematic with these methods due to its considerably large size, its huge repetitive content and its complicated as well as complex transcriptional network [2]. On the other hand, Nandi et al. [7] have introduced the randomized occurrence frequency ( $OF_r$ ) as the average number of positive predictions in the random shuffled promoter sequences and determined muscle specific TFs which occur together with the transcription factor MyoD within a certain distance of 100bp. Hu et al. [17] have used in their work the hypergeometric test to identify synergistic TF interactions in tissue specific genes. While the approach of Nandi et al. mainly takes into account tissue specific properties of interacting TFs, the approach of Hu et al. principally considers the enriched TFBS combinations in overlapping orthologous genes of human and mouse which leads to ignoring the detection of non-enriched but interacting TF-pairs. Further, these methods require user specified parameters such as the level of significance of the test performed or a background random set which is likely to affect their performance.

Recently, a novel method called MatrixCatch has been introduced by Deyneko et al. [6] to identify CRMs in promoter sequences. Mainly focusing on the experimentally verified CRMs, MatrixCatch recognizes in individual sequences the known TF pairs from the TRANSCompel<sup>®</sup> [19] database. Although this method significantly outperforms several statistical methods, it clearly disregards the pairs which are not included in TRANSCompel<sup>®</sup>. As a result of this, MatrixCatch reaches an improved performance in identifying CRMs with a significantly higher nucleotide-level correlation coefficient (nCC) value in comparison to other methods, but it is not able to detect novel TF pairs which can be also crucial for understanding gene regulation.

In this study, we propose a method called Potentially Collaborating Transcription Factor Finder (PC-TraFF) to detect interactions between homotypic and heterotypic transcription factor pairs using pointwise mutual information (PMI). PMI is a very useful association measure in the field of linguistics for document summarization processes as well as for the detection of combinations of words in a corpus indicating that those words have some idiosyncrasy in their linguistic distribution [20–23]. We adopt the PMI in the field of bioinformatics replacing words in a document with TFBSs in a set of sequences to develop our new method, which includes following main steps. First, we replace the Term-Sentence-Matrix, suggested by Aji S et al. [20] for document summarization, with a TFBS-Sequence-Matrix (TSM) to characterize the importance of each TFBSs in a sequence with respect to the entire set of sequences. Thereafter, according to a predefined distance between TFBSs, PC-TraFF builds all possible TFBS-pairs and calculates their weighted pointwise mutual information scores. Unlike previous methods [6–8, 17], PC-TraFF estimates for each TFBS pair the expected levels of background PMI arising from the random noise of false positive TFBSs using the average product correction (APC) suggested by Dunn et al. [24]. Finally, the weighted PMI values of each TFBS pair are corrected by the APC theorem.

The aim of this study is to identify collaborating TFs that frequently bind in a cooperative manner in a set of genomic sequences. Our results show that a large majority of significant pairs found by PC-TraFF in promoter sequences of different RefSeq genes and miRNA genes are in agreement with previous experimental studies. In addition to finding biologically characterized TF pairs, PC-TraFF is able to identify additional potentially collaborating TFs which could provide new targets for future works.

## Results

In this study, we introduce PC-TraFF, a computational method that aims to identify potential collaborating

transcription factors based on their binding sites. Our method comprises the following steps. For a given set of sequences, we first determine the transcription factor binding sites (TFBSs) applying the Match™ program [25] with vertebrate position weight matrices (PWMs) from TRANSFAC [26]. Second, we construct a TFBS-sequence matrix to display the occurrence of unique TFBSs in each sequence and then filter this matrix in order to eliminate highly over- and/or underrepresented TFBSs in all sequences. Third, by calculating the pointwise mutual information (PMI) between each sequence and each TFBS in the filtered TFBS-sequence matrix, we identify the important TFBSs indicating that they occur in the corresponding sequences more than by chance. Afterwards, considering these important TFBSs in our further analysis, we build TFBS pairs based on predefined minimal and maximal distances between their coordinates on the DNA. Next, the weighted cumulative pointwise mutual information  $\text{PMI}_{pc}$  between TFBSs of a pair is calculated to define their collaboration level in the entire set of sequences. Employing the average product correction (APC) theorem [24] to reduce the background noise due to false positive TFBSs, we correct the  $\text{PMI}_{pc}$ -values of TFBS pairs. Finally, transforming the corrected  $\text{PMI}_{pc}$ -values into z-scores, we define a pair to be significant if it has a z-score  $\geq 3$ .

The Results section of this work comprises three parts. First, to investigate the performance of PC-TraFF we made a pairwise comparison with the previous methods MatrixCatch [6], CPModule [9], and CrmMiner [27]. Second, to further test the functionality of PC-TraFF significant TFBS pairs we performed for human promoters of RefSeq genes and miRNA genes: i) a genome-wide gene set analysis where each promoter region is represented by the 1000 bp upstream of the TSS of all annotated genes; ii) a breast cancer subtype-associated gene set analysis whose promoter regions are defined by Joshi et al. [28] as 500 bp upstream to 100 bp downstream relative to the corresponding TSSs. Third, we present the computational time and memory consumption of PC-TraFF in comparison to MatrixCatch [6], CPModule [9], and CrmMiner [27].

As a prerequisite for our approach, we had to define for the TFBSs in a pair minimal distance and maximal distance constrains. However, we only demonstrate in this section results for minimal distance  $\geq 5$ , maximal distance  $\leq 20$ . The remaining results can be found in Additional file 1.

After predicting PC-TraFF significant TFBS pairs in the corresponding set of sequences, we validate those pairs mainly focusing on the TRANSCompel® (release 2014.2) [19], BioGRID interaction database (version 3.2.119) [29] and STRING database [30] since all of them contain experimentally proven pairs. Further literature search is done if we cannot validate a pair in those databases.

### Comparisons with existing methods

To investigate the state-of-the-art prediction quality of pointwise mutual information measure proposed in this work, we were interested to determine the overlap between the TFBS pairs predicted by different methods. Thus we made pairwise comparisons between our new PC-TraFF, MatrixCatch [6], CPModule [9], and CrmMiner [27]. For this comparison study, we applied PC-TraFF using different distance measures. It is important to note that we only selected the methods which are applicable to the human genome and the software implementation of which is ready-to-use. All four methods take as input a sequence set and a PWM library satisfying certain admissibility criteria. As a result, PC-TraFF, CPModule, and CrmMiner output a set of significant TFBS pairs, but MatrixCatch outputs all predicted pairs without any significance threshold for a sequence set. To make MatrixCatch results comparable with the results of these three methods, we determined the frequency of each pair in MatrixCatch outcomes and then took the top ranking pairs whose frequencies are equal or bigger than average. Further, there is a fundamental difference between these methods: while PC-TraFF and MatrixCatch do not require any background set, to apply CPModule and CrmMiner a background set is needed.

The results of this comparison are threefold. First, we applied these methods to the promoter sequences of RefSeq genes in the genome-wide analysis as well as the breast cancer analysis to determine the overlap of their predictions. Second, we randomly selected 200 promoter sequences (-1000 bp relative to the TSSs) from chromosome 21, hence it has in average similar GC content to human genome. In these 200 sequences, we inserted the TFBS pair (V\$IRF1\_01 - V\$USF1\_01) which represents the interaction between transcription factors IRF1 and USF1. The minimal and maximal distances between these TFBSs are defined as at least 5 bp and at most 20 bp, respectively. Further, the TFBS pair was sampled in each sequence between two to twelve times, randomly (see Additional file 2). Third, we computed the sensitivity, specificity, and Matthews correlation coefficient (MCC) values to assess the performance of PC-TraFF and the three previous methods.

Let  $\mathcal{N}_{\text{PC-TraFF}} := (\mathcal{V}_{\text{PC-TraFF}}, \mathcal{E}_{\text{PC-TraFF}})$  denote the predicted collaboration network of TFBS pairs where any two elements of  $\mathcal{N}_{\text{PC-TraFF}}$  are connected by an undirected edge belonging to  $\mathcal{E}_{\text{PC-TraFF}}$  if and only if the corresponding TFBS pair is PC-TraFF significant. By extending this concept in full analogy, we observed for each of these methods the predicted collaboration networks  $\mathcal{N}_{\text{PC-TraFF}_{20}}$ ,  $\mathcal{N}_{\text{pctff}_{50}}$ ,  $\mathcal{N}_{\text{PC-TraFF}_{100}}$ ,  $\mathcal{N}_{\text{MC}}$ ,  $\mathcal{N}_{\text{CPM}}$ , and  $\mathcal{N}_{\text{CrmM}}$ , where  $\mathcal{N}_{\text{PC-TraFF}_{20,50,100}}$  indicate the application of PC-TraFF with different distance measures and  $\text{MC}$ ,

CPM, CrmM stand for the abbreviation of MatrixCatch, CPModule, and CrmMiner, respectively.

First, we performed the overlap comparison between methods edge-oriented using the number of overlapping edges as measure. Applying these methods to the sequences of RefSeq genes in the genome-wide analysis and breast cancer analysis, the number of predicted TFBS pairs as well as the number of overlapping pairs is calculated as  $|\mathcal{E}_{PC-TraFF_{20}}|$ ,  $|\mathcal{E}_{PC-TraFF_{50}}|$ ,  $|\mathcal{E}_{PC-TraFF_{100}}|$ ,  $|\mathcal{E}_{MC}|$ ,  $|\mathcal{E}_{CPM}|$ ,  $|\mathcal{E}_{CrmM}|$ ,  $|\mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{PC-TraFF_{50}}|$ ,  $|\mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{PC-TraFF_{100}}|$ ,  $|\mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{PC-TraFF_{100}}|$ ,  $|\mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{MC}|$ ,  $|\mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{CPM}|$ ,  $|\mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{CrmM}|$ ,  $|\mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{MC}|$ ,  $|\mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{CPM}|$ ,  $|\mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{CrmM}|$ ,  $|\mathcal{E}_{PC-TraFF_{100}} \cap \mathcal{E}_{MC}|$ ,  $|\mathcal{E}_{PC-TraFF_{100}} \cap \mathcal{E}_{CPM}|$ , and  $|\mathcal{E}_{PC-TraFF_{100}} \cap \mathcal{E}_{CrmM}|$ , which are displayed in Tables 1 and 2.

Although all methods perform a combinatorial search of frequently occurring TFBS pairs and aim to identify their significance in the given set of sequences, Table 1 shows that each of these methods detects in the same set of sequences using the same PWM library considerably different numbers of important TFBS pairs. The reason for that can be explained due to the differences in their underlying algorithms. While MatrixCatch mainly scans the sequences to recognize the known pairs from TransCompel database, CPModule applies a very stringent TFBS screening threshold with an additional filtering step based on nucleosome occupancy, which results in a dramatic reduction of significant pairs found by CPModule. On the other hand, CrmMiner uses a supervised classification approach for the identification of significantly enriched TFBS pairs in the sequences under study.

Table 2 suggests that regardless of the distance measure used, a large amount of TFBS pairs are regularly detected by PC-TraFF as significant. Further, Table 2 clearly demonstrates that all of these methods carry distinct information and thus the overlap between any two of them is quite low. Thus the pairwise comparison highly indicates that under the assumption that each of these methods focuses on different important aspects of interaction between TFs, they can complement each other perfectly. Especially, this assumption is true for PC-TraFF as an information theory-based method compared with the other three conventional methods.

**Table 2** Total number of edges in two predicted collaboration networks of different methods

	Total number of common edges in collaboration networks	
	Genome-wide analysis	Breast cancer analysis
$ \mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{PC-TraFF_{50}} $	43	54
$ \mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{PC-TraFF_{100}} $	41	43
$ \mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{MC} $	3	1
$ \mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{CPM} $	6	0
$ \mathcal{E}_{PC-TraFF_{20}} \cap \mathcal{E}_{CrmM} $	0	0
$ \mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{PC-TraFF_{100}} $	82	80
$ \mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{MC} $	4	1
$ \mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{CPM} $	8	1
$ \mathcal{E}_{PC-TraFF_{50}} \cap \mathcal{E}_{CrmM} $	2	0
$ \mathcal{E}_{PC-TraFF_{100}} \cap \mathcal{E}_{MC} $	4	1
$ \mathcal{E}_{PC-TraFF_{100}} \cap \mathcal{E}_{CPM} $	9	0
$ \mathcal{E}_{PC-TraFF_{100}} \cap \mathcal{E}_{CrmM} $	2	0
$ \mathcal{E}_{MC} \cap \mathcal{E}_{CPM} $	1	0
$ \mathcal{E}_{MC} \cap \mathcal{E}_{CrmM} $	0	1
$ \mathcal{E}_{CPM} \cap \mathcal{E}_{CrmM} $	3	1

Second, we applied all of these methods to the randomly selected sequence set, explained above. While PC-TraFF and CPModule successfully detected the inserted TFBS pair as significant, MatrixCatch and CrmMiner have not detected this pair.

To assess the performance of PC-TraFF, we further made a statistical comparison between our method and the three previous methods. For this comparison study, we followed a similar procedure suggested by Yu et al [31]. As positive controls we obtained in total 3158 TFBS pairs according to experimentally validated interactions between TFs from TRANSCompel<sup>®</sup>, BioGRID and STRING interaction databases. As negative controls, we used all possible remaining pairs which have not been experimentally validated yet but could be predicted based on the PWM library applied in this study. Having applied all methods to the above mentioned promoter sequences, we observed that each of these methods reaches considerably high specificity and quite low sensitivity indicating that all methods show comparable performances. The details are presented in Table 3. As expected, all methods suffer from low sensitivity because the way how we assess this parameter is a very tough one, leading to a

**Table 1** Total number of edges in method-dependent significant collaboration networks

Sequence sets of RefSeq genes in	Total number of edges in predicted collaboration network					
	$ \mathcal{E}_{PC-TraFF_{20}} $	$ \mathcal{E}_{PC-TraFF_{50}} $	$ \mathcal{E}_{PC-TraFF_{100}} $	$ \mathcal{E}_{MC} $	$ \mathcal{E}_{CPM} $	$ \mathcal{E}_{CrmM} $
Genome-wide analysis	54	86	91	19	17	21
Breast cancer analysis	64	82	88	13	6	25



**Table 3** Performance comparison between PC-TraFF<sub>20</sub>, PC-TraFF<sub>50</sub>, PC-TraFF<sub>100</sub>, MatrixCatch (MC), CPModule (CPM), and CrmMiner (CrmM)

	Sensitivity	Specificity	MCC
PC-TraFF <sub>20</sub>	2.3 %	99.5 %	0.088
PC-TraFF <sub>50</sub>	3.1 %	99.3 %	0.10
PC-TraFF <sub>100</sub>	3.2 %	99.3 %	0.102
MC	0.5 %	99.9 %	0.053
CPM	0.5 %	100 %	0.06
CrmM	0.6 %	99.6 %	0.025

large overestimation of false negatives. Thus, the consideration of sensitivity alone is of limited value and should be taken for comparison of the different methods only. Further, our results indicate that the usage of PC-TraFF with different distance constrains gives rise to prediction of different numbers of TFBS pairs (see Table 1) which slightly changes its performance (see Table 3). Considering MCC-values, our PC-TraFF reaches moderately increased performance compared to the three other methods. Thus, we propose mutual usage of previous methods with PC-TraFF together so that they can complement each other (for details see Table 4).

**Table 4** The complementary usage of different methods can lead to an improved performance in identifying important pairs in sequences

	Sensitivity	Specificity	MCC
PC-TraFF <sub>20</sub> U MC	2.8 %	99.5 %	0.101
PC-TraFF <sub>50</sub> U MC	3.6 %	99.3 %	0.112
PC-TraFF <sub>100</sub> U MC	3.8 %	99.3 %	0.114
PC-TraFF <sub>20</sub> U CPM	2.6 %	99.5 %	0.099
PC-TraFF <sub>50</sub> U CPM	3.4 %	99.3 %	0.107
PC-TraFF <sub>100</sub> U CPM	3.5 %	99.3 %	0.109
PC-TraFF <sub>20</sub> U CrmM	3.0 %	99.2 %	0.087
PC-TraFF <sub>50</sub> U CrmM	3.8 %	99 %	0.10
PC-TraFF <sub>100</sub> U CrmM	3.9 %	99 %	0.102
MC U CPM	1.0 %	99.9 %	0.079
MC U CrmM	1.2 %	99.6 %	0.050
CPM U CrmM	1.2 %	99.6 %	0.051
PC-TraFF <sub>20</sub> U MC U CPM	3.1 %	99.5 %	0.11
PC-TraFF <sub>50</sub> U MC U CPM	3.8 %	99.3 %	0.118
PC-TraFF <sub>100</sub> U MC U CPM	4 %	99.3 %	0.12
PC-TraFF <sub>20</sub> U MC U CPM U CrmM	3.8 %	99.2 %	0.10
PC-TraFF <sub>50</sub> U MC U CPM U CrmM	4.5 %	99 %	0.116
PC-TraFF <sub>100</sub> U MC U CPM U CrmM	4.7 %	99 %	0.119
MC U CPM U CrmM	1.7 %	99.6 %	0.07

Additionally, we compared the predictions of PC-TraFF, MatrixCatch, CPModule, and CrmMiner, which have not been experimentally validated yet. It turned out that there is only one TFBS pair (V\$MYC\_MAX\_B - V\$EGR\_Q6) that is experimentally unconfirmed, but even so, detected by PC-TraFF and CrmMiner as significant.

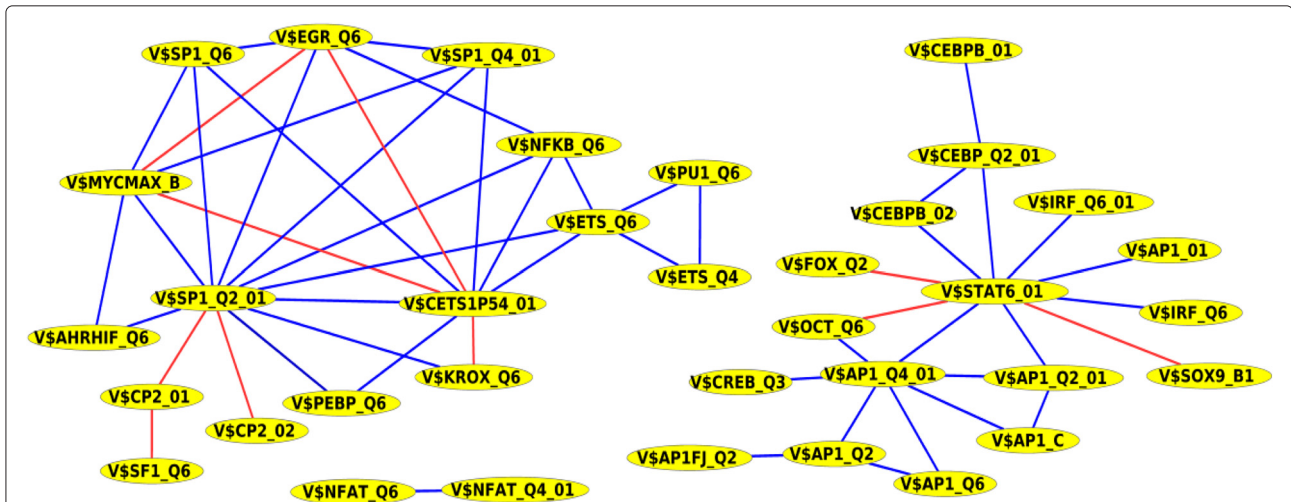
#### A genome-wide analysis of promoters in the context of RefSeq genes and miRNA genes

Applying our method to 23015 promoter sequences of human RefSeq genes, we observed 54 PC-TraFF significant collaborating TFBS pairs which are comprised of 7 homotypic and 47 heterotypic pairs. According to their z-scores, the top 10 PC-TraFF significant pairs determined in promoter sequences of human RefSeq genes are given in Table 5 (for the whole list of significant pairs see Additional file 3). The importance of 44 pairs out of all significant pairs has been experimentally verified by previous studies regarding their interactions which are summarized in TRANSCompe<sup>®</sup> [19], BioGRID [29] and STRING [30] interaction databases. The remaining 10 TFBS pairs found by PC-TraFF have not been experimentally validated yet and the reason for their significance is still unclear.

As shown in Fig. 1, the predicted collaboration network of PC-TraF significant TFBS pairs is comprised of three unconnected subgraphs and consists of 35 nodes and 54 edges where each edge refers to a collaboration and each node corresponds to a TFBS. Moreover, the network contains the four hubs V\$SP1\_Q2\_01,

**Table 5** Significant TFBS pairs found by PC-TraF in genome-wide promoter analysis of human RefSeq genes. The table shows the top 10 significant TFBS pairs, which are sorted in descending order based on their z-scores

Significant pair	Z-score	Reference
V\$PU1_Q6 - V\$ETS_Q6	9.84	TRANSCompe <sup>®</sup> , BioGRID, STRING
V\$CETS1P54_01 - V\$ETS_Q6	5.76	TRANSCompe <sup>®</sup> , BioGRID, STRING
V\$ETS_Q4 - V\$ETS_Q6	5.49	TRANSCompe <sup>®</sup> , BioGRID, STRING
V\$EGR_Q6 - V\$SP1_Q2_01	5.09	BioGRID, STRING
V\$CETS1P54_01 - V\$SP1_Q2_01	4.94	TRANSCompe <sup>®</sup> , STRING
V\$AP1_Q2_01 - V\$AP1_Q4_01	4.69	TRANSCompe <sup>®</sup> , BioGRID
V\$STAT6_01 - V\$OCT_Q6	4.66	-
V\$CEBPB_02 - V\$STAT6_01	4.58	TRANSCompe <sup>®</sup> , STRING
V\$MYC_MAX_B - V\$SP1_Q2_01	4.36	BioGRID, STRING
V\$AP1FJ_Q2 - V\$AP1_Q2	4.09	TRANSCompe <sup>®</sup> , BioGRID, STRING



**Fig. 1** PC-TraFF significant collaborating TFBS pairs based on promoter sequences of human RefSeq genes. Blue lines denote interactions between TFs whose importance is experimentally verified whereas red lines indicate potential interactions between transcription factors that have not been experimentally validated yet

V\$STAT6\_01, V\$CETS1P54\_01, and V\$AP1\_Q4\_01 each of which provides critical knowledge to understand mechanisms of the gene regulatory network. The hubs and their top three collaboration partners are given in Table 6.

The binding site V\$SP1\_Q2\_01 is a GC-rich motif on the DNA bound by Sp1 which is a member of the three-zinc finger Krüppel-related transcription factors family [32]. Initially, Sp1 was detected as a general TF needed for the activation of a large number of housekeeping genes. In addition, Sp1 is important for the recruitment of the transcriptional machinery in the absence of a TATA box [33, 34]. Sp1 interacts with corepressors or coactivators to regulate transcription in cell-signaling events and

to modulate DNA-binding specificity [35, 36]. The second hub in the network is the binding site V\$STAT6\_01 bound by the factor STAT6 belonging to the family of STAT factors which seldomly activate transcription alone but act together with other factors to active transcription [37–39]. STAT6 is known to be involved in the immune system. Here, it acts in response to the cytokines IL-4 and IL-13 and thus it is required for T-cell proliferation as well as responses in T-cells [40]. In addition, STAT6 was recently identified to function in non-immune tissues like mammary gland, lung and skin [40]. Another hub is V\$CETS1P54\_01 representing the binding site of ETS1 which is a member of the evolutionarily conserved ETS family of transcription factors [41, 42]. The factor

**Table 6** The hubs and their top three collaboration partners in the predicted collaboration network of significant TFBS pairs for human RefSeq genes

Hub	Top three collaborating pairs	Z-score	Reference
V\$SP1_Q2_01	V\$EGR_Q6	5.09	BioGRID, STRING
	V\$CETS1P54_01	4.94	TRANSCompel®, STRING
	V\$MYCMAX_B	4.36	BioGRID, STRING
V\$STAT6_01	V\$OCT_Q6	4.66	-
	V\$CEBPB_02	4.58	TRANSCompel®, STRING
	V\$CEBP_Q2_01	3.74	TRANSCompel®, BioGRID, STRING
V\$CETS1P54_01	V\$ETS_Q6	5.76	TRANSCompel®, BioGRID, STRING
	V\$SP1_Q2_01	4.94	TRANSCompel®, STRING
	V\$NFKB_Q6	3.96	TRANSCompel®, STRING
V\$AP1_Q4_01	V\$AP1_Q2_01	4.69	TRANSCompel®, BioGRID, STRING
	V\$STAT6_01	3.35	TRANSCompel®, BioGRID, STRING
	V\$AP1_Q6	3.35	TRANSCompel®, BioGRID, STRING

ETS1 plays a critical role in T-cell and B-cell proliferation and differentiation [41, 43]. Moreover, ETS1 is one of the well investigated transcription factors whose transcriptional activity is regulated by other factors by physical and functional interactions [41, 44, 45]. The next hub in the network is V\$AP1\_Q4\_01 which is bound by AP-1 transcription factor. Simplified, AP-1 is a heterodimer of JUN and FOS proteins or a homodimer of JUN proteins. All AP-1 constituents belong to the leucine zipper family, known as the one of the largest family of dimerizing TFs in humans that share as a common feature a bZIP domain [1, 32, 46, 47]. There is a huge number of different AP-1 proteins which are all differentially expressed and regulated indicating that the dimers differ in their cellular function [48]. In general, AP-1 is involved in cell proliferation and differentiation as well as cell cycle progression. Its combinatorial interactions with other transcription factors are required for the specification of (regulatory) transcriptional activities of FOS-JUN family proteins in the human genome [48–50].

A closer look at the predicted collaboration network of significant TFBS pairs (see Fig. 1) and Table 6 reveals that the hub TFBS pairs V\$SP1\_Q2\_01 - V\$CETS1P54\_01 bound by Sp1 - ETS1 and V\$STAT6\_01 - V\$AP1\_Q4\_01 bound by STAT6 - AP-1 (JUN) exhibit significant cooperativity in their binding. The interaction between Sp1 and ETS1 appears among others in TATA-less promoters where the TATA-box can be replaced by a non-consensus binding site for Sp1. The binding of Sp1 to this site is of low affinity, but can be strengthened by the interaction to ETS1 bound adjacent to it on DNA [51]. The physical interaction between STAT6 and JUN was observed to play a critical role in the upregulation of the IL-24 promoter. IL-24 is a multifunctional cytokine that is important for B cell differentiation as well as anticancer effects in diverse cancer cells [52].

Above, we concentrated our research on interactions of TFs with RefSeq genes. To extend our knowledge about the gene regulatory network, we will in the following also address the question of TF-miRNA gene interactions. However, it is important to note that promoters of miRNA genes used in this study are based on the predicted TSSs. Consequently, they should not be treated as reliable as the TSSs of RefSeq genes and the results may somewhat vary when working with the results of different prediction algorithms. It has been demonstrated that TFs can regulate miRNAs as well as miRNAs can regulate TFs. Additionally, both are involved in gene regulation, TFs on a transcriptional level, miRNAs on a translational one. It might therefore be interesting to compare the transcriptional networks for genes and miRNAs regarding interacting TFs to find similarities or dissimilarities. For this purpose, we further performed a genome-wide analysis with PC-TraFF of the promoters of human miRNAs

using computationally predicted promoter sequences of miRNAs over ca. 50 tissues and cell lines (see Additional file 4). Applying PC-TraFF to these human miRNA promoters, we observed 42 significant TFBS pairs, among which 35 heterotypic and 7 homotypic pairs could be identified. The top 10 PC-TraFF significant pairs determined in promoter sequences of human miRNA genes are given in Table 7 (for the whole list of significant pairs see Additional file 5).

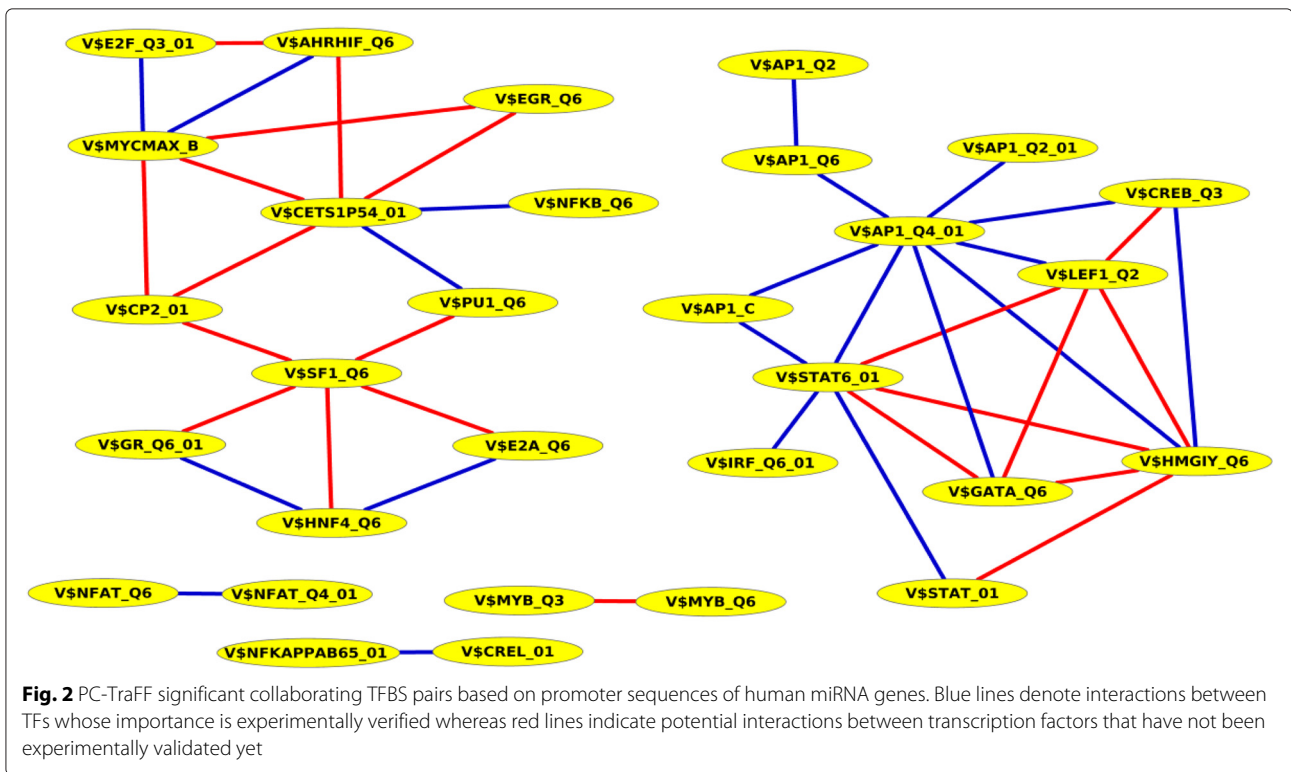
In addition, 14 of 42 significant TFBS pairs overlap with the result of promoter sequence analysis of human RefSeq genes. The importance and functionality of these significant pairs was checked with the TRANSCOMPel® [19]. BioGRID [29] and STRING interaction databases [30]. Here, biological importance of 21 TFBS pairs could be confirmed through interaction databases. The remaining 21 PC-TraFF significant TFBS pairs have not been experimentally validated yet and the reason for their significance is still unclear.

Like the TFBS pair analysis of human RefSeq genes, we constructed based on the significant TFBS pairs found by PC-TraFF of human miRNA promoters a predicted collaboration network. It consists of 30 nodes and 42 edges where each edge refers to a collaboration and each node corresponds to a TFBS (see Fig. 2). The most remarkable result of this analysis is that the network contains the three hubs V\$AP1\_Q4\_01, V\$CETS1P54\_01, and V\$STAT6\_01 which have been also identified as hubs in the significant TFBS pairs collaboration network of human RefSeq genes (see Fig. 1). The hubs and their top three collaboration partners are given in Table 8.

Previous studies described that AP-1, which binds to the V\$AP1\_Q4\_01 motif, is involved in the expression of several miRNAs. For example, AP-1 activates miR-155 in the

**Table 7** Significant TFBS pairs found by PC-TraFF in genome-wide promoter analysis of human miRNA genes. The table shows the top 10 significant TFBS pairs, which are sorted in descending order based on their z-scores

Significant pair		Z-score	Reference
V\$STAT6_01	- V\$HMGY_Q6	13.73	-
V\$HMGY_Q6	- V\$LEF_Q2	5.89	-
V\$HMGY_Q6	- V\$GATA_Q6	5.18	-
V\$CREB_Q3	- V\$AP1_Q4_01	5.16	BioGRID, STRING
V\$MYC_MAX_B	- V\$AHRIF_Q6	5.03	BioGRID, STRING
V\$STAT6_01	- V\$AP1_Q4_01	4.98	TRANSCOMPel®, BioGRID, STRING
V\$HMGY_Q6	- V\$AP1_Q4_01	4.97	BioGRID, STRING
V\$STAT6_01	- V\$LEF_Q2	4.83	-
V\$SF1_Q6	- V\$HNF4_Q6	4.79	-
V\$HMGY_Q6	- V\$CREB_Q3	4.79	BioGRID, STRING



processes of B-cell activation and maturation [53]. ETS1 binds to the V\$CETS1P54\_01 motif and regulates among others the expression of miR-126, which is responsible for the regulation of angiogenesis and vascular inflammation [54]. STAT6 binds to V\$STAT6\_01 and is involved in the cholesterol biosynthesis pathway through targeting miR-197 [55]. Besides this, it has been described to be regulated by miRNAs which act among others as tumor suppressors [56].

Furthermore, it is important to note that the hub TFBSs V\$STAT6\_01 and V\$AP1\_Q4\_01 were detected by PC-TraFF as a significant pair indicating that their bindings frequently occur in a cooperative manner in the promoter

sequences of human miRNA like in the promoters of human RefSeq genes.

**Analysis of breast cancer subtype-associated promoter regions**

Today, it is widely known that breast cancer is the most common cancer in women. Breast cancer can be separated into five subgroups termed Luminal A, Luminal B, Normal-like, ErbB2 over-expressing and Basal-like [28]. In order to expand our analysis to more specific, clinically relevant situations, we applied our new method to promoter regions of breast cancer-associated RefSeq genes and their regulating miRNA genes.

**Table 8** The hubs and their top three cooperation pairs in the predicted collaboration network of significant TFBS pairs for human miRNA genes

Hub	Top three collaborating pairs	Z-score	Reference
V\$AP1_Q4_01	V\$CREB_Q3	5.16	BioGRID, STRING
	V\$STAT6_01	4.98	TRANSCompel®, BioGRID, STRING
	V\$HMGY_Q6	4.97	BioGRID, STRING
V\$CETS1P54_01	V\$MYCMAX_B	4.33	-
	V\$PU1_Q6	3.67	TRANSCompel®, BioGRID, STRING
	V\$EGR_Q6	3.64	-
V\$STAT6_01	V\$HMGY_Q6	13.73	-
	V\$AP1_Q4_01	4.98	TRANSCompel®, BioGRID
	V\$LEF_Q2	4.82	-

Similar to the genome-wide analysis, we started with analyzing the 218 promoter regions of target RefSeq genes. As a result of this analysis, we observed 64 PC-TraFF significant collaborating TFBS pairs that are comprised of five homotypic and 59 heterotypic pairs (see Additional file 6). The biological importance of 44 pairs has been experimentally verified by previous studies whereas the remaining 20 PC-TraFF significant pairs have not been experimentally validated yet and the reason for their significance is still unclear.

Interestingly, we found that two TFBSs in the PC-TraFF significant pairs are representing the E2F transcription factor family (see Fig. 3). In general, this family is known to be involved in cell cycle regulation as well as apoptosis and DNA damage response. Our results reveal that members of the E2F family are collaborating with each other which has been proven by experimental studies in the context of breast cancer [57]. Briefly, activating and repressive E2Fs bind to adjacent sites on the BRCA1 promoter and regulate its activity. In response to hypoxia, they cause the downregulation of unmutated BRCA1 which in turn is associated with sporadic cancers of the breast [57]. In our study, we further detected the established collaboration of E2F family members with Sp1, c-Myc and NF- $\kappa$ B1, each of which plays a critical role in breast cancer [34, 58, 59]. The interaction of E2F and Sp1 has been experimentally verified to play a fundamental role in the activation of S-phase specific promoters at the G<sub>1</sub>/S boundary of the cell cycle [60].

The binding site V\$NFKB\_Q6 that is bound by members of the NF- $\kappa$ B related factors family forms a hub in the network of potential collaborating pairs of the breast cancer gene set (see Fig. 3 and Table 9). In general, NF- $\kappa$ B related factors are involved in the regulation of cell processes like proliferation, survival and immunity. In addition, they are critical for the regulation of inflammation as well as angiogenesis [61] and are known to be involved in breast cancer [59]. In our study, we found that NF- $\kappa$ B1,

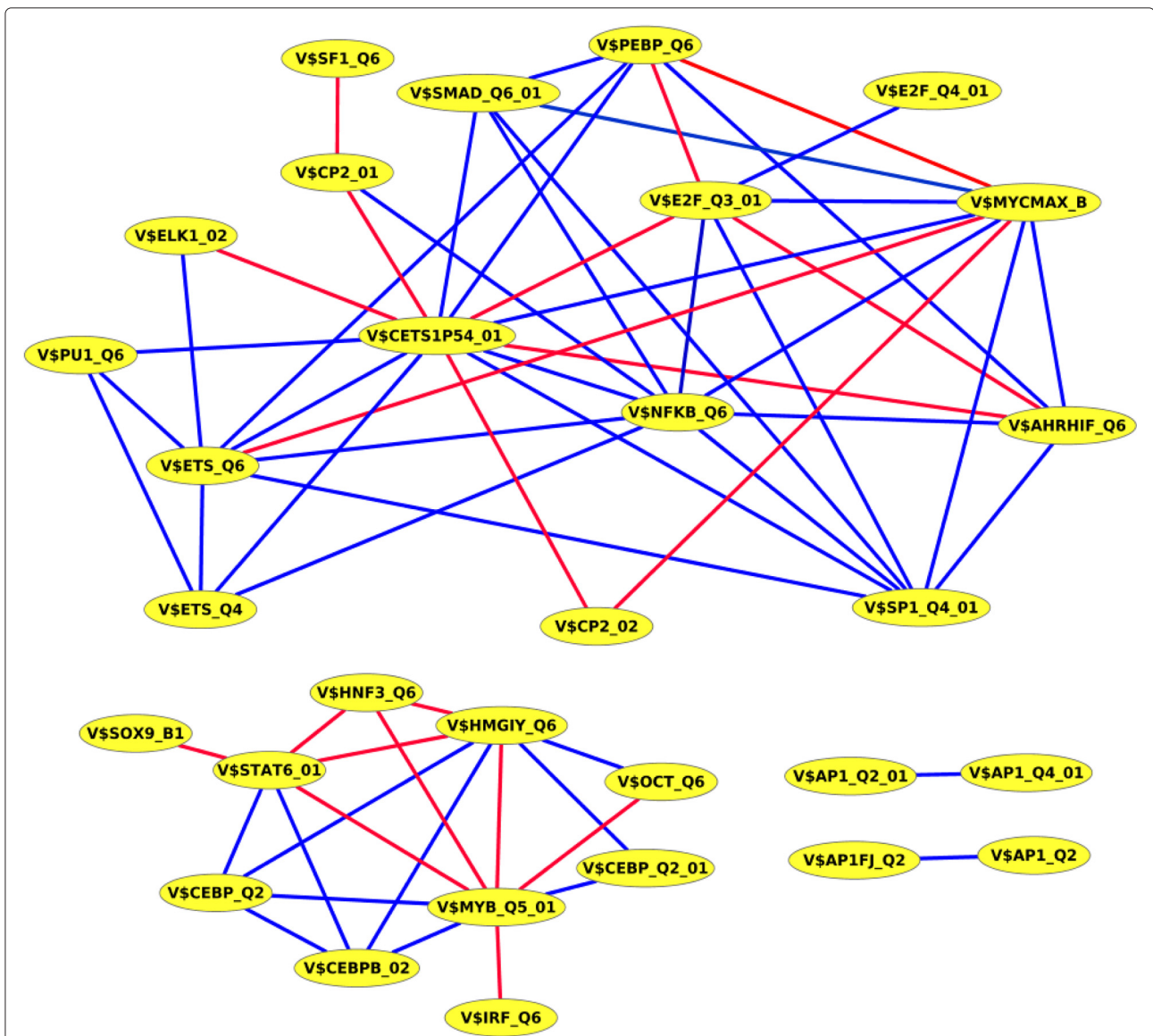
a member of the family NF- $\kappa$ B related factors [32], interacts with ETS1, ELF1, Sp1, and E2F1. ETS1 is involved in breast cancer where it regulates genes that are important for metastasis and tumor progression [62]. ELF1 belongs to the Ets-related factors family and regulates genes that are involved in cell growth and differentiation. Its overexpression is linked with breast cancer [63]. Another member of the NF- $\kappa$ B related factors family is RelA which is found to collaborate with SMAD3, AHR and c-Myc each of which is known to be involved in breast cancer [64, 65]. AHR is a ligand activated transcription factor whose activity is linked with alterations in cell proliferation, apoptosis, adipose differentiation, tumor promotion, immune function, vitamin A status, development and reproductive functions [66]. The physical interaction of RelA and AHR is important for the activation of the c-Myc oncogene in breast cancer cells [65].

Three TFBSs in our significant pairs (V\$CEBP\_Q2, V\$CEBPB\_02 and V\$CEBP\_Q2\_01) can be bound by transcription factor C/EBP $\beta$ . This TF is known to regulate genes that are involved in invasion, cellular proliferation, survival and apoptosis [67]. Further, the level of C/EBP $\beta$  is often increased in metastatic breast cancer and is known to correlate with a high tumor grade [67]. We found this factor interacting with HMGA1, c-Myb and STAT6. HMGA1 is regulating gene expression by altering the chromatin structure and orchestrating transcription factor complexes to enhanceosomes within promoter regions [68]. Additionally, it is known to be overexpressed in aggressive cancers and to be involved in metastatic progression in triple negative breast cancers [68]. The interaction of HMGA1 and C/EBP $\beta$  is in particular crucial for the regulation of the human insulin receptor [69]. c-Myb functions in cell differentiation as well as cell proliferation and is involved in different types of tumors [70].

To gain more insight into the role of TF interactions in gene regulatory networks, we further applied PC-TraFF to the promoters of breast cancer-associated miRNAs. In our analysis, we found 43 PC-TraFF significant collaborating

**Table 9** The hubs and their top three collaboration partners in the predicted collaboration network of breast cancer-associated significant TFBS pairs for human RefSeq genes

Hub	Top three collaborating pairs	Z-score	Reference
V\$NFKB_Q6	V\$CETS1P54_01	5.42	TRANSCompel <sup>®</sup> , STRING
	V\$ETS_Q6	4.80	BioGRID, TRANSCompel <sup>®</sup> , STRING
	V\$SP1_Q4_01	3.43	BioGRID, TRANSCompel <sup>®</sup> , STRING
V\$CETS1P54_01	V\$ETS_Q6	8.01	BioGRID, TRANSCompel <sup>®</sup> , STRING
	V\$NFKB_Q6	5.42	TRANSCompel <sup>®</sup> , STRING
	V\$MYCMAX_B	5.21	-
V\$MYCMAX_B	V\$CETS1P54_01	5.16	-
	V\$E2F_Q3_01	5.21	TRANSCompel <sup>®</sup>
	V\$AHRHIF_Q6	4.39	BioGRID, STRING

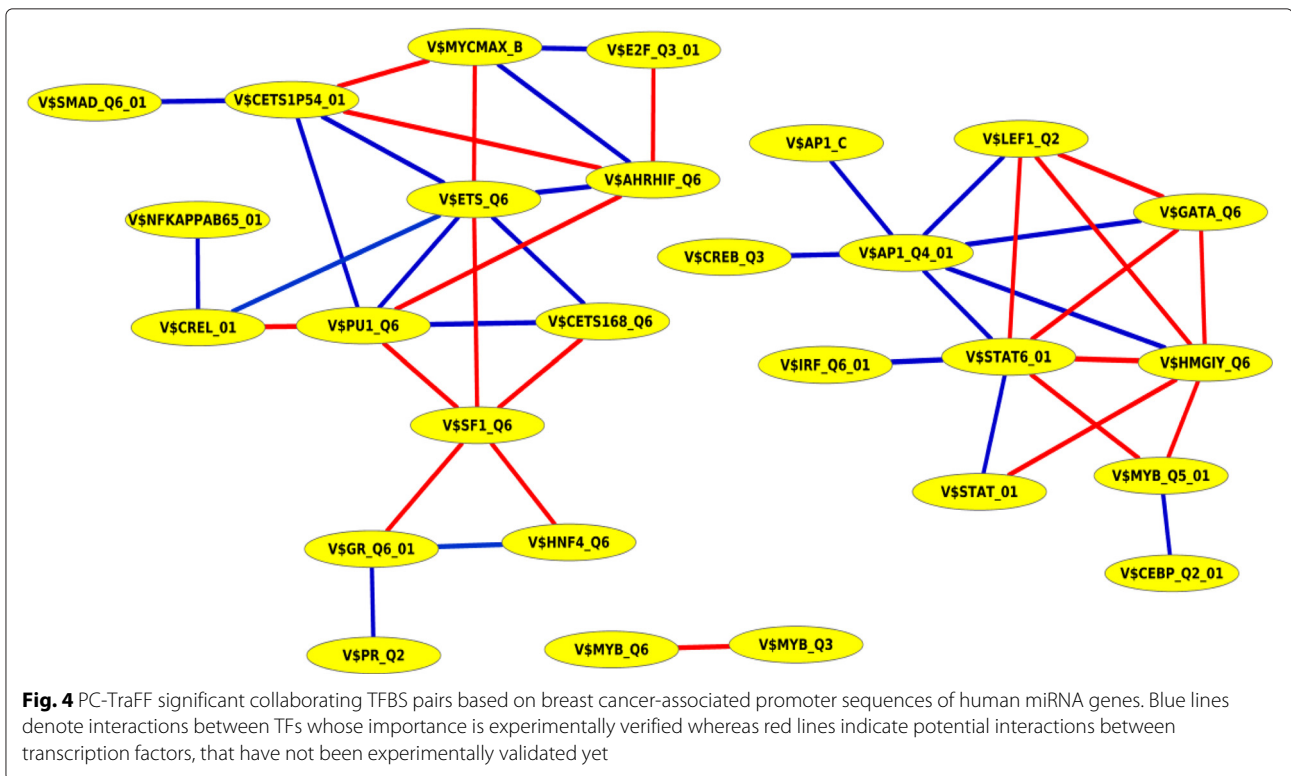


**Fig. 3** PC-TraFF significant collaborating TFBS pairs based on breast cancer-associated promoter sequences of human RefSeq genes. Blue lines denote interactions between TFs whose importance is experimentally verified whereas red lines indicate potential interactions between transcription factors that have not been experimentally validated yet. The binding sites V\$NFKB\_Q6, V\$CETS1P54\_01, and V\$MYC\_MAX\_B constitute three hubs in the predicted collaboration network of significant TFBS pairs. The hubs and their top three collaboration partners are given in Table 9

TFBS pairs that are comprised of 8 homotypic and 35 heterotypic pairs (see Fig. 4). 14 out of 43 significant pairs have been also detected by PC-TraFF in the breast cancer-associated promoters of RefSeq genes. Of all significant pairs 22 could be verified based on annotation databases TransCompel, BioGRID and/or STRING. The significance of the remaining pairs is still unclear. In addition to interactions between TFs in the promoters of miRNA genes, we further investigated the interplay between TFs and miRNAs. Consequently, we found for TFs in 37 pairs at least a reference to their interaction with miRNAs in literature (see Additional file 7).

Figure 4 shows that the collaboration network contains the five hubs V\$STAT6\_01, V\$SETS\_Q6, V\$AP1\_Q4\_01, V\$HMGIIY\_Q6, and V\$PUI\_Q6 each of which plays a critical role in the breast cancer-associated gene regulatory network [62, 68, 71–74]. The hubs and their top three collaboration partners are given in Table 10. V\$SETS\_Q6 is bound by ETS1 which also binds to V\$CETS1P54\_01 and V\$CETS168\_Q6. Both are found to collaborate with V\$SETS\_Q6 and show quite high significance levels in the PC-TraFF analysis. ETS1 has been described in literature to be involved in regulation of and by miRNAs which are involved in cancer [54, 75]. As an example, ETS1 has





been found to regulate and is in turn also regulated by miR-222 [75]. It was found that a phosphorylated part of the ETS1 protein induced miR-222 transcription in metastatic melanoma [75]. As previously described, ETS1 is additionally involved in regulation of miR-126 [54]. This miRNA is also known to be involved in breast cancer

regulation, more specifically, it has been observed to act as a metastasis suppressor miRNA in human breast cancer [76]. The transcription factor PU.1 binds to sites predicted with V\$PU1 Q6. It has been shown to be important for differentiation and development of several cell types and tissues, as for example in B cell development and

**Table 10** The hubs and their top three collaboration partners in the predicted collaboration network of significant TFBS pairs for breast cancer-associated human miRNA genes

Hub	Top three collaborating pairs	Z-score	Reference
V\$STAT6_01	V\$HMGY_Q6	13.28	-
	V\$MYB_Q5_01	5.77	-
	V\$GATA_Q6	4.98	-
V\$ETS_Q6	V\$PU1_Q6	13.49	TRANSCOMPEL®, BioGRID, STRING
	V\$SF1_Q6	6.16	-
	V\$CETS1P54_01	5.00	TRANSCOMPEL®, BioGRID, STRING
V\$AP1_Q4_01	V\$HMGY_Q6	4.85	BioGRID, STRING
	V\$LEF1_Q2	4.27	BioGRID
	V\$STAT6_01	4.17	TRANSCOMPEL, BioGRID, STRING
V\$HMGY_Q6	V\$STAT6_01	13.28	-
	V\$MYB_Q5_01	6.17	-
	V\$LEF1_Q2	6.00	-
V\$PU1_Q6	V\$ETS_Q6	13.49	TRANSCOMPEL®, STRING
	V\$SF1_Q6	5.88	-
	V\$CETS168_Q6	3.29	TRANSCOMPEL®, BioGRID, STRING

terminal myeloid differentiation [77]. Additionally, it has been described to be associated with cancer, as it interacts with the p53 family of tumor suppressors and acts as a tumor suppressor itself in B cell malignancies [77, 78]. Like ETS1, PU.1 is involved in miRNA regulation and has been reported to regulate the transcription of miR-142 in hematopoietic cell specific expression as well as miR-424 expression in human monocyte and macrophage differentiation [79, 80]. Another hub is V\$AP1\_Q4\_01, which is bound by AP-1. This TF has been shown to be involved in regulation of miR-21, a miRNA which has been observed to be significantly deregulated in breast cancer [81, 82].

### Comparative analysis of breast cancer subtypes

Breast cancer tumors can be separated into five different subgroups with unique RefSeq genes based on their mRNA expression patterns. As has been noted in [28], the promoters of the individual subtypes can be distinguished by their composition of TFBS. The number of promoter sequences of RefSeq genes as well as the corresponding number of PC-TraFF significant pairs found for each subtype is shown in Table 11. The results show that there is a certain pairwise overlap between the significant pairs found in all subtypes (see Table 12) indicating that some TF collaborations are not restricted to the individual subtypes. The largest pairwise overlap with 36 significant pairs is between Luminal A and Luminal B indicating that this subtypes match in a large part of their regulatory features. There is further a huge significant TFBS pair overlap found in Luminal A and Basal-like as well as Luminal B and Basal-like associated sequences.

Six significant pairs (see Table 13) are detected by PC-TraFF in all subtypes, each of them has been detected as significant previously (see Fig. 3). One of these pairs represents the synergistic collaboration between transcription factors PEBP2 $\alpha$ A and ETS1 whose direct interaction is crucial for the activation of the osteopontin (Opn) promoter [83]. Opn is in general important for ossification [83] but its splicing variants have been shown to be expressed in breast cancer cells [84]. Another TFBS pair out of these six pairs represents the collaboration between

**Table 11** Number of promoter sequences of breast cancer subtype-associated RefSeq genes and corresponding significant pairs found by PC-TraFF

Subtype	Number of sequences	Number of Pairs
Luminal A	86	61
Luminal B	57	62
Basal-like	31	72
Normal-like	27	49
ErbB2 over-expressing	16	62

**Table 12** Number of pairwise overlapping significant pairs of the RefSeq genes of breast cancer subtypes Luminal A, Luminal B, Basal-like, Normal-like, and ErbB2 over-expressing

Subtype	Luminal A	Luminal B	Basal-like	Normal-like	ErbB2 over-exp.
Luminal A	-	36	28	26	23
Luminal B		-	30	20	19
Basal-like			-	25	19
Normal-like				-	16
ErbB2 over-exp.					-

C/EBP $\beta$  and STAT6 which often bind directly adjacent on DNA and activate transcription in a synergistic manner [85].

In analogy to our previous analysis, we investigated in the next step the interactions between TFs in the promoter sequences of breast cancer subtype-associated miRNA genes. The number of promoter sequences of miRNA genes as well as the number of PC-TraFF significant pairs identified for each subtype is shown in Table 14. As for the breast cancer subtype-associated Refseq genes, we made a pairwise overlap comparison between the significant pairs identified in the promoters of subtype-associated miRNA genes (see Table 15). Similar to the previous findings, the results of this comparison show that the largest pairwise overlap is found between Luminal A and Luminal B with 38 overlapping pairs whereas the smallest significant TFBS pair overlap is found between the Basal-like and the ErbB2 over-expressing subtype. Further the results suggest that the significant TFBS pairs found in each subtypes do not vary clearly. In contrast to the Refseq gene analysis, in the miRNA promoters 20 PC-TraFF significant TFBS pairs have been detected in all five subtypes (see Table 16). Surprisingly, one of these pairs, namely V\$SF1\_Q6 and V\$E2A\_Q6 does not occur in the predicted TFBS pair collaboration network of miRNA genes of the breast cancer analysis (see Fig. 4). The binding sites V\$SF1\_Q6 and V\$E2A\_Q6 are bound by the factors NR5A2 and TCF3, respectively. NR5A2 has been described to be associated with invasive breast cancer and

**Table 13** Six PC-TraFF significant TFBS pairs found in promoter sequences of RefSeq genes of all five breast cancer subtypes

Significant pairs	Reference
V\$MYCMAX_B - V\$E2F_Q3_01	TRANSCompel <sup>®</sup>
V\$CETS1P54_01 - V\$PEBP_Q6	TRANSCompel <sup>®</sup> , BioGRID, STRING
V\$CETS1P54_01 - V\$NFKB_Q6	TRANSCompel <sup>®</sup> , STRING
V\$CEBP_Q2 - V\$STAT6_01	TRANSCompel <sup>®</sup> , BioGRID, STRING
V\$AP1_Q2_01 - V\$AP1_Q4_01	TRANSCompel <sup>®</sup> , BioGRID, STRING
V\$CEBPB_02 - V\$STAT6_01	TRANSCompel <sup>®</sup> , STRING



**Table 14** Number of breast cancer subtype-associated miRNA genes and corresponding significant pairs found by PC-TraFF

Subtype	Number of miRNAs	Number of Pairs
Luminal A	186	46
Luminal B	53	61
Basal-like	76	45
Normal-like	23	52
ErbB2 over-expressing	70	45

is additionally thought to be involved in promotion of migration of breast cancer [86]. TCF3 upregulates miR-495 in breast cancer stem cells [87]. Additionally, TCF3 is supposed to be involved in breast cancer growth and initiation and is preferentially highly expressed in breast cancer with poor prognosis of the basal-like subtype [88]. Although both transcription factors are involved in breast cancer, we could not confirm their direct interaction through annotation databases or literature survey.

**Computational time and memory usage of PC-TraFF**

The identification of significant TFBS pairs in human genome is computationally intensive because of its considerably large size and its complicated as well as complex transcriptional network. When analysing a set of sequences of the human genome, the computational time and memory usage can rise very quickly due to the huge number of potential TFBS pairs. Thus, one of our main targets while developing PC-TraFF algorithm was to keep its computational time and memory usage tractable. PC-TraFF is implemented in Java and performed on Intel Core™ i7-4770K Processor operating at 3.50 GHz, with 32 GB DDR3 RAM using Ubuntu 12.04.5 operating system (64 - bit version). Further, we compared the performance of PC-TraFF with MatrixCatch [6], CPMModule [9], CrmMiner [27], CisMiner [4], and COPS [8]. However, our attempt to apply CisMiner and COPS to human genomic sequences failed because the scope of applicability of both methods is strongly limited due to their very high execution time and memory consumption.

**Table 15** Number of pairwise overlapping significant pairs of the miRNA analysis of breast cancer subtypes Luminal A, Luminal B, Basal-like, Normal-like, and ErbB2 over-expressing

Subtype	Luminal A	Luminal B	Basal-like	Normal like	ErbB2 over-exp.
Luminal A	-	38	28	31	30
Luminal B	-	-	31	32	33
Basal-like	-	-	-	27	24
Normal-like	-	-	-	-	27
ErbB2 over-exp.	-	-	-	-	-

**Table 16** 20 PC-TraFF significant TFBS pairs found in promoter sequences of miRNA genes of all five breast cancer subtypes

Significant pairs		Reference
V\$STAT6_Q1	-	V\$HMGY_Q6 -
V\$HMGY_Q6	-	V\$L1_Q2 -
V\$HMGY_Q6	-	V\$MYB_Q5_Q1 -
V\$STAT6_Q1	-	V\$MYB_Q5_Q1 -
V\$SF1_Q6	-	V\$CETS168_Q6 -
V\$HMGY_Q6	-	V\$AP1_Q4_Q1 BioGRID, STRING
V\$STAT6_Q1	-	V\$AP1_Q4_Q1 TRANSCmpel®, BioGRID, STRING
V\$STAT6_Q1	-	V\$GATA_Q6 -
V\$HMGY_Q6	-	V\$GATA_Q6 -
V\$GATA_Q6	-	V\$L1_Q2 -
V\$MYC_MAX_B	-	V\$AHRHIF_Q6 BioGRID, STRING
V\$AP1_C	-	V\$AP1_Q4_Q1 TRANSCmpel®, BioGRID, STRING
V\$SF1_Q6	-	V\$E2A_Q6 -
V\$SF1_Q6	-	V\$HNF4_Q6 -
V\$GATA_Q6	-	V\$AP1_Q4_Q1 TRANSCmpel®, BioGRID, STRING
V\$L1_Q2	-	V\$AP1_Q4_Q1 BioGRID
V\$MYC_MAX_B	-	V\$E2F_Q3_Q1 TRANSCmpel®
V\$NFKAPPAB65_Q1	-	V\$CREL_Q1 BioGRID, STRING
V\$STAT_Q1	-	V\$HMGY_Q6 -
V\$E2F_Q3_Q1	-	V\$AHRHIF_Q6 -

Applying PC-TraFF algorithm to the promoter sequences of RefSeq genes, the average computational time of a sequence was 0.1806 s in genome-wide promoter analysis and 0.0203 s in breast cancer analysis, respectively. Consequently, the algorithm took ~ 69 minutes with a memory requirement of 3229 Mb for genome-wide analysis and less than one minute (~ 0.07 minute) with a memory requirement of 581 Mb for breast cancer analysis. The computational time and memory usage of PC-TraFF in comparison to other tools is presented in Table 17.

**Table 17** Computational time (in seconds) / memory usage (in megabyte) of the individual tools

	Genome-wide analysis	Breast cancer analysis
PC-TraFF	4158.4 s / 3229 Mb	4.4 s / 581 Mb
CPModule	2213.0 s / 721.6 Mb	5.9 s / 7.8 Mb
CrmMiner	34409.6 s / 526 Mb	857.4 s / 90 Mb
MatrixCatch	627.2 s / 70.7 Mb	16.9 s / 46.2 Mb

## Discussion

Previous studies showed that Pointwise Mutual Information (PMI) is a powerful association measure in the field of linguistics. Aji S et al. [20] used PMI in their study for document summarization processes based on a Term-Sentence-Matrix where they measured weights of words to describe their importance in sentences. On the other hand, Gerlof Bouma [21] applied PMI in his work for extracting collocations from a text where he aimed to identify essential word combinations in sentences which display some idiosyncrasy in their linguistic distributions. These two articles encouraged us to utilize PMI for the identification of potentially collaborating transcription factors based on the idiosyncrasy of their binding site distributions on the genome. Thus adopting the idea of Aji S et al. [20] and Gerlof Bouma [21] in the field of bioinformatics, we treat in this study the genome as a document, the sequences under investigation as sentences, and TFBSs as words in these sentences.

Today, it is known that in higher organisms TFs often form non-random combinations of functional dimers or higher order complexes instead of acting alone. Until now, different studies have confirmed that the binding sites of TFs provide a useful clue in the prediction of collaborating TFs in a set of sequences (see e.g. [4–14]). As a result, we use the TFBSs as the key components of PC-TraFF. However in our method the challenge was to filter these TFBSs with the objective of eliminating the bias as well as noise effects of both highly over- and underrepresented TFBSs in a consistent way. These highly over- and underrepresented TFBSs could be assumed to be punctuation marks or stop words like “a”, “the”, “of” etc. which are required in sentences due to the grammatical structures of natural languages. However they do not provide meaningful information in statistical analysis for the identification of important words in sentences [20]. Moreover, we apply an additional filtering step in order to avoid the overestimation of such TFBS pairs which directly overlap with TFBSs of their same type (see the “Methods” section, Phase 3). These overlaps result from the palindromic TFBSs and the PWMs used by Match<sup>®</sup> program [25]. The filtering can be seen as removal of redundant words in sentences indicating that these words do not contribute any additional information about the content of a sentence.

Another fundamental step of our new method is the construction of TFBS pairs for which a distance measure between TFBSs according to their localization is required. Today, different approaches are utilized to define the distance constraints between TFBSs like the calculation of the preferred distances between TFBSs based on their coordinates on the sequences (see e.g. [4, 8]) or the usage of certain predefined maximum and minimum distances between TFBSs (see e.g. [11, 17, 27]). As suggested by Hu et al. [11], in this study we preferred

the latter approach and tested our method using different predefined distance constraints. However our distance definition between TFBSs clearly differs from the previous definitions used in [8, 11], hence in these studies the distance between TFBSs has been calculated based on the last nucleotide of the first TFBS and first nucleotide of the second TFBS. We find the usage of this definition doubtful in our study since: i) it can result in negative distances if we consider slightly overlapping TFBSs which satisfy our predefined maximum and minimum distance constraints; ii) we believe that the first or last nucleotide of a TFBS is not convincing since the borders of TFBSs as they are represented by PWMs are somewhat fuzzy.

In order to almost completely eliminate the noise of false positive TFBSs, we additionally applied the average product correction (APC) theorem. The APC theorem is a promising method which has been developed by Dunn et al. [24] as an explicit noise measure based on information theory to estimate the background mutual information of residue positions in multiple sequence alignments. This theorem seems to be of universal applicability and thus we utilized it in our approach to calculate for each TFBS pair the background  $PMI_{pc}(t_a; t_b)$  shared by TFBSs  $t_a$  and  $t_b$  in the set of sequences under study. By removal of the background from the observed  $PMI_{pc}$ -values, the pointwise mutual information is decreased which results in the correction of the observed values. As a consequence, a separation of the signal caused by functional collaboration of TFs from the background occurs. We use these corrected values for ranking the candidate pairs without influence of noise contained in the sequences under study.

The results we present in this study for different sets of sequences of human RefSeq genes show that the vast majority of TFBS pairs found by PC-TraFF are in agreement with previous experimental studies. 44 significant TFBS pairs in the genome-wide analysis of promoters as well as in the breast cancer-associated sequence set analysis, respectively, have been confirmed by literature regarding to the interactions of corresponding TFs. Such interactions contribute crucial information for our understanding of combinatorial aspects of gene regulatory networks in the human cell cycle [2]. To gain more insights into the regulatory network we further analyzed the promoter regions of miRNA genes whose interactions with TFs play an important role in several biological processes [89]. Unlike recent studies [89–92], which mainly focus on the interplay between miRNAs and single TFs, in our analysis we systematically studied the interactions between TFs in the promoters of miRNA genes. It turned out that there are several overlapping significant pairs which are detected in the sequences of both miRNA genes and RefSeq genes indicating that the collaboration of corresponding TFs are essential for transcription in general. However, we found one binding site V\$HMGIIY\_Q6

which was found more frequently in the significant TFBS pairs in the promoters of miRNA genes than RefSeq genes. V\$HMG1Y\_Q6 is bound by the transcription factors HMGA1 and HMGA2. Mammalian HMGA proteins have been shown to play key roles in chromatin architecture and gene control and are known to have oncogenic activity [93]. Furthermore, it has been shown that HMGA proteins regulate miRNAs. For example, the miRNAs miR-196a-2, miR-101b, miR-331 and miR-29a have been found to be downregulated in cells lacking the HMGA1 protein [93]. Additionally, the miRNA miR-181b has been shown to be up-regulated by HMGA1 and both are supposed to be involved in breast cancer progression [94]. This, in correlation with our results, might hint to the fact that the HMGA proteins could be important regulators of miRNAs.

Of particular interest, we created based on the PC-TraFF significant TFBS pairs for each analysis a collaboration network (see Figs. 1, 2, 3 and 4). These networks support us on the one hand for explaining the potential biological functions of TF pairs in the corresponding set of sequences. On the other hand, they help us to generate new hypotheses for extending our knowledge of why these transcription factors tend to bind in a preferential manner. All collaboration networks of significant pairs contain two large unconnected subgraphs. These findings are consistent with those of Hu et al. [11] and indicate that the collaboration networks of transcription factors are split in two major groups according to their binding behaviour. Interestingly, we explore that the predicted collaboration networks for RefSeq genes as well as miRNA genes in the genome-wide analysis contain the binding sites V\$STAT6\_01, V\$CETS1P54\_01, and V\$AP1\_Q4\_01 with a higher degree of connectivity and thus they are defined as hubs in both networks. However, the binding site V\$SP1\_Q2\_01 shows a sole exception in the genome-wide analysis in comparison to other hubs because we can only find it in the collaboration network for RefSeq genes. The reason why this binding site can not form a significant pair in the genome-wide analysis of miRNA genes, is still unclear. For the breast cancer-associated sequence set analysis, the predicted collaboration networks for miRNA genes and their target RefSeq genes contain completely different binding sites as hubs. This finding indicates that the functional interactions between TFs for the regulation of the miRNA transcription could also differ from the interactions between TFs for the gene regulation of RefSeq genes. We further analyzed breast cancer subtype specific sets of sequences by separating the breast cancer-associated sequences into five subgroups as has been noted in [28]. A comparison between the significant pairs found in all subtypes reveals that PC-TraFF detected six experimentally verified TFBS pairs (see Table 13) which are found and are likely to play a critical role in each

subtype. The results further suggest that our method is not dependent on the number of sequences under study, since the PC-TraFF can detect for a small number of sequences a high number of significant TFBS pairs or vice versa.

Additionally, we applied the PC-TraFF using different distance constraints as suggested by Hu et al. [11]. The results denote that a considerable number of true significant TFBS pairs are consistently detected by PC-TraFF under different distance constraints which indicates the consistency of PC-TraFF predictions (see Additional file 1).

Although we can verify the importance of most TFBS pair predictions in the promoter regions of human RefSeq genes, there are still 10 and 20 unconfirmed TFBS pairs found for the genome-wide analysis and breast cancer-associated sequence set analysis, respectively. It is interesting to note that three of the unconfirmed TFBS pairs (V\$CETS1P54\_01 – V\$MYC\_MAX\_B, V\$CP2\_01 – V\$SF1\_Q6, and V\$SOX9\_B1 – V\$STAT6\_01) are referred as significant in both analyses. As discussed in [31], one reason for the significant co-occurrence of all unconfirmed binding sites could be that their TFs do not have direct physical interaction but rather collaborate with each other through another co-factor indirectly. However, we hypothesize that most of the unconfirmed pairs identified by our present method in the promoter regions of both RefSeq genes as well as miRNA genes may play a critical role for an effective regulation of the transcriptional machinery in both analysis notwithstanding the absence of previous experimental data. Therefore, further progress from the biochemistry and molecular biology end is required not only to evaluate the significance of these pairs, but also for a future perspective on a deeper understanding of regulatory networks.

Finally, we made a pairwise comparison between the results of PC-TraFF and conventional methods Matrix-Catch [6], CPModule [9], and CrmMiner [27]. This comparison study reveals that all these methods detect remarkably different sets of TFBS pairs as important which results in considerably low overlaps between the results of all these methods. The reason for that can be explained that all methods model different aspects of interactions between transcription factors and thus carry distinct information. However, the comparison results additionally indicate that all these methods reach comparable performances. These findings are consistent with those of Klepper et al. [95] where they applied several methods to identify TFBS pairs using different datasets and then showed that no single method is better than other. Thus, we suggest to use these methods together to improve the performance in identifying important pairs.

## Conclusions

In this study, we develop PC-TraFF for the identification of potentially collaborations between TFs using their binding site distributions on the sequences under study. PC-TraFF is a new information theoretic method that applies the pointwise mutual information by considering TFBSs like words and sequences like sentences. PC-TraFF also utilizes the average product correction theorem which reduces the effect of false positive TFBSs and thus enhances the signal caused by functional interactions between TFs. Results show that PC-TraFF algorithm has a tractable computational time and memory consumption. Our results further indicate that PC-TraFF is on the one hand able to identify known collaborating pairs in the sequences, on the other hand able to predict additional pairs which are likely to play critical role in the gene regulatory network but have not been experimentally validated yet. Thus we suggest that the web server of PC-TraFF could be used as a novel automated tool for the prediction of potential collaborating transcription factors which are required to better understand the molecular mechanism of cellular regulation.

## Methods

### Set of sequences for RefSeq genes and miRNA genes

Using UCSC genome browser [96], we obtain for human RefSeq genes and miRNA genes the corresponding promoter sequences based on their annotated transcription start sites (TSS). It is important to note that while the TSSs of RefSeq genes have been obtained from the UCSC genome browser, the TSSs of miRNA genes have been determined during an internal project, the publication of which is under preparation. The method utilized for obtaining the TSS of the miRNAs depends on the positions of modified histones, more precisely the positions of H3K4me3. This modified histone has been described to be localized mainly at the promoters and TSS of transcriptionally active genes in the genome [97]. Therefore, these positions in collaboration with some computational TSS identifying tools were used to define the TSS and promoter regions of miRNAs. Moreover, it is important to note that we have also analysed the promoter sequences of miRNAs from PROmiRNA database [98] to compare its results to those of our data. It turned out that there are several overlapping significant pairs found by PC-TraFF (data not shown).

In this study, the assembly of the hg19 release of the human genome was used and only UCSC track refGene annotations were considered whose chromosome annotations correspond to the chromosomes chr1–chr22, chrX and chrY.

Regarding TSS annotations, RefSeq genes and miRNA genes can have highly correlated multiple promoters which results in overestimation of some transcription

factor binding sites (TFBSs). Thus, to avoid the redundancy between sequences we filter them based on their TSSs and use in our analysis only those sequences which have no overlap.

### TFBS detection

We scan each sequence and its reverse complement employing the Match™ program [25] setting its profile parameter as specified by Deyneko et al. in [6] to detect transcription factor binding sites (TFBSs). To apply the Match™ program, we used a vertebrate position weight matrix (PWM) library suggested in [6]. The PWMs were obtained from the latest version of TRANSFAC (release 2014.1) [26].

### The PC-TraFF algorithm

The PC-TraFF algorithm consists of six phases to detect potentially collaborating transcription factors in a set of sequences.

#### Phase 1: construction and filtering of the TFBS-sequence matrix

Based on the frequency of predicted TFBSs in each sequence, we create a TFBS-sequence matrix  $\mathbb{M}$ , where rows correspond to IDs of the sequences and columns refer to names of PWMs. The entries of  $\mathbb{M}$  are calculated as follows. Let  $s_i$  ( $i = 1, \dots, m$ , where  $m$  is the number of sequences) denote a promoter sequence and let  $t_j$  ( $j = 1, \dots, n$ , where  $n$  is the number of PWMs under study) be a potential TFBS predicted by PWM  $j$ . The entry of  $\mathbb{M}$  at position  $(i, j)$ ,  $f_{ij}$ , is calculated as the observed frequency of  $t_j$  in the sequence  $s_i$ .

Afterwards, we filter  $\mathbb{M}$  in order to reduce: i) the bias of the highly represented TFBSs in all sequences; ii) the noisy effect of false signals arising from insufficient data. Hence, we define for a matrix  $\mathbb{M}$  its filtering parameters as follows. First, we calculate the standard deviation  $\sigma$  of the entire matrix  $\mathbb{M}$  based on its column sums. After that, we eliminate a column  $k$  in  $\mathbb{M}$  if the column sum of  $k$  is greater than  $3 \times \sigma$ . Second, we identify average zero percentile in  $\mathbb{M}$  based on its column entries and remove all columns in  $\mathbb{M}$  if such columns consist of more zero entries than average, as we formally received the best results with this approach.

#### Phase 2: identification of important TFBSs in each sequence

Using the filtered matrix  $\mathbb{M}$ , the importance of each TFBS in each sequence is characterized by calculating the pointwise mutual information between sequence  $s_i$  and TFBS  $t_j$  ( $\text{PMI}_{st}$ ) as

$$\text{PMI}(s_i; t_j) = \log_2 \frac{p(s_i, t_j)}{p(s_i) \cdot p(t_j)}, \quad (1)$$

where  $p(s_i, t_j)$  indicates the probability that TFBS  $t_j$  occurs in the sequence  $s_i$  with respect to entire set of sequences. It is calculated as

$$p(s_i, t_j) = \frac{f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}}, \tag{2}$$

where  $f_{ij}$  is the frequency of the TFBS  $t_j$  in the corresponding sequence  $s_i$ .

$p(s_i)$  and  $p(t_j)$  are the marginal probabilities for  $s_i$  and  $t_j$  in the entire set of sequences, respectively, which are calculated as

$$p(s_i) = \frac{\sum_{j=1}^n f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}}, \tag{3}$$

$$p(t_j) = \frac{\sum_{i=1}^m f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}}. \tag{4}$$

A positive  $\text{PMII}(s_i; t_j)$ -score for a specific TFBS  $t_j$  in the sequence  $s_i$ , resulting from the fact that the pair distribution  $p(s_i, t_j)$  is greater than the product of the marginal distributions, shows that  $t_j$  occurs in  $s_i$  more often than by chance. Conclusively, we regard such TFBSs in sequences as important for transcription and consider only those TFBSs in our further analysis for each sequence.

**Phase 3: filter to avoid overlaps**

The Match™ program predicts all potential TFBSs based on the given PWM library. Thereby, it is possible that some binding sites overlap or one binding site is included in another. The overlap between binding sites can occur due to: i) the palindromicity of TFBSs (the reverse complement is the same as the original sequence); ii) some PWMs being larger than real binding sites of TFs.

Overlapping of TFBSs of the same type can result in their overestimation in our analysis. Thus, to avoid the overestimation of such TFBSs, we filter them based on their distance to the corresponding TSS. After the filtering process, the TFBS is taken into account that has a closer distance to TSS compared to its overlapping partner (illustrated in Fig. 5) since functional TFBSs often have a closer localization to TSSs [37].

**Phase 4: construction of TFBS pairs**

We define the distance,  $d_{t_A, t_B}$  between two TFBSs  $t_A$  and  $t_B$  based on their midpoints  $C_{t_A}$  and  $C_{t_B}$ :

$$d_{t_A, t_B} = |C_{t_A} - C_{t_B}| \tag{5}$$

The midpoint,  $C_{t_A}$  of a TFBS  $t_A$  is defined as  $\lfloor \frac{\text{length}_{t_A}}{2} \rfloor$  where  $\text{length}_{t_A}$  is the length of  $t_A$ .

In this work, two TFBSs form a pair, if  $d_{min} \leq d_{t_A, t_B} \leq d_{max}$  where  $d_{min}$  and  $d_{max}$  are minimal and maximal distance constrains, respectively, which are specified by user. In this study, we set  $d_{min}$  at least 5 bp which approximately corresponds to one-half of an average TFBS' length. In analogy to study of Hu et al. [11], we used different  $d_{max}$  constrains in our analysis. Moreover, following [99] a slight overlap (of at most 4 bp) between TFBSs of different types is allowed if the user-defined distance constrains are satisfied.

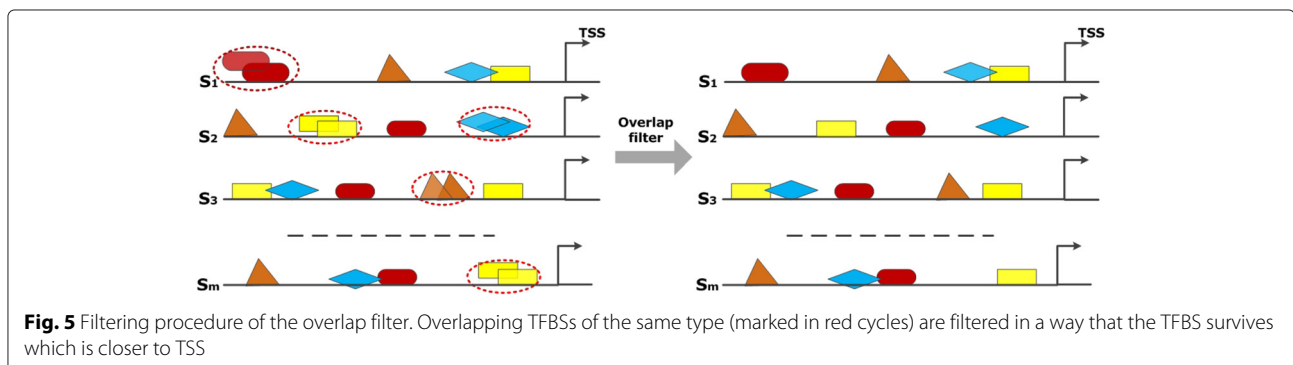
Applying our approach to construct TFBS pairs, we have to deal with their false overestimation due to repeated number of similar binding sites within a certain interval on DNA, also known as homotypic clustering. To avoid this problem in our analysis, we allow that one TFBS can only participate in a pair of two specified TFBSs within a certain interval (predefined distance). This is illustrated in Fig. 6.

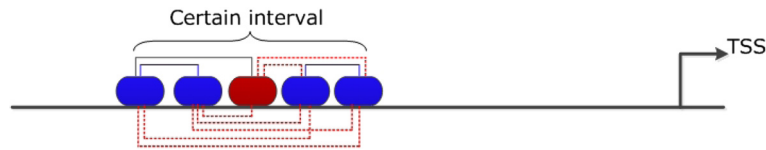
**Phase 5: weighted cumulative pointwise mutual information**

Potential collaborating transcription factors are determined by calculating weighted cumulative pointwise mutual information ( $\text{PMII}_{pc}$ ) based on the co-occurrences of their corresponding TFBSs. The  $\text{PMII}(t_a; t_b)$  between TFBSs  $t_a$  and  $t_b$  is defined as

$$\text{PMII}(t_a; t_b) = \log_2 \frac{p(t_a, t_b)}{p(t_a) \cdot p(t_b)}, \tag{6}$$

where  $p(t_a, t_b)$  is the joint probability,  $p(t_a)$  and  $p(t_b)$  are marginal probabilities for  $t_a$  and  $t_b$ , respectively. In general, the  $\text{PMII}$ -metric is very susceptible to low number counts [21]. To eliminate this property of the  $\text{PMII}$ -metric to some extent, we first multiply the  $\text{PMII}(t_a; t_b)$ -value of each TFBS pair with their joint probability  $p(t_a, t_b)$ . After





**Fig. 6** The problem of homotypic clusters: The TFBSs ( $t_{blue}$ ) form an homotypic cluster within a certain interval on the sequence. The TFBS  $t_{red}$  is also included in this interval. According to our definition to construct TFBS pairs and by following the DNA strand in 5'-3' direction: i) we consider one  $t_{blue} - t_{red}$  pair in this interval indicating that an individual TFBS can only participate in one count of a specified pair (shown with black line); ii) if we consider  $t_{blue} - t_{blue}$  pairs, there are two pairs within this interval (shown with blue lines). The red (dashed) lines demonstrate that the remaining  $t_{blue} - t_{blue}$  and  $t_{blue} - t_{red}$  pairs are not taken into account in the calculation of pointwise mutual information of this pairs

that, we incorporate the weight of each sequence ( $w_s$ ) with respect to the entire set of sequences in the calculation of PMI. Doing this, the weighted pointwise mutual information of each TFBS pair in a sequence  $s$   $\text{PMII}_p^s(t_a; t_b)$  is obtained as

$$\text{PMII}_p^s(t_a; t_b) = w_s \cdot p(t_a, t_b) \cdot \text{PMI}(t_a; t_b). \quad (7)$$

The sequence weight  $w_s$  for a sequence  $s$  is given by the number of TFBS pairs  $N_s$  in  $s$  divided by the total number of TFBS pairs in the entire set of sequences  $S$ .

$$w_s = \frac{N_s}{\sum_{s_i \in S} N_{s_i}} \quad (8)$$

To define the collaboration level of  $t_a$  and  $t_b$  in  $S$ , we calculate weighted cumulative pointwise mutual information value  $\text{PMII}_{pc}(t_a; t_b)$  by summing up their  $\text{PMII}_p^s(t_a; t_b)$ -values over all sequences as

$$\text{PMII}_{pc}(t_a; t_b) = \sum_{s \in S} \text{PMII}_p^s(t_a; t_b). \quad (9)$$

#### Phase 6: background noise reduction of TFBSs using average product correction

We apply the average product correction (APC) procedure, developed by Dunn *et al.* [24], to reduce the background noise of TFBS pairs that might occur as a result of false positive TFBSs in the entire sequence set  $S$ . Thus, we estimate the expected level of the background  $\text{PMII}_{pc}(t_a; t_b)$  shared by TFBSs  $t_a$  and  $t_b$  as

$$\text{APC}(t_a, t_b) = \frac{\overline{\text{PMII}_{pc}(t_a; \bar{t}_x)} \cdot \overline{\text{PMII}_{pc}(t_b; \bar{t}_x)}}{\overline{\text{PMII}_{pc}}}, \quad (10)$$

where  $\overline{\text{PMII}_{pc}(t_a; \bar{t}_x)}$  is the mean pointwise mutual information of TFBS  $t_a$  that is defined by

$$\overline{\text{PMII}_{pc}(t_a; \bar{t}_x)} = \frac{1}{n-1} \sum_{x=1}^n \text{PMII}_{pc}(t_a; t_x). \quad (11)$$

Further, the  $\overline{\text{PMII}_{pc}}$  refers to overall mean pointwise mutual information for all TFBS pairs.

Afterwards, the  $\text{APC}(t_a, t_b)$ -value of a pair under study is subtracted from its  $\text{PMII}_{pc}(t_a; t_b)$ -value, and thus we

observe the corrected  $\text{PMII}_{pc}^{APC}(t_a; t_b)$ -values as

$$\text{PMII}_{pc}^{APC}(t_a; t_b) = \text{PMII}_{pc}(t_a; t_b) - \text{APC}(t_a, t_b) \quad (12)$$

Finally, by transforming the corrected  $\text{PMII}_{pc}^{APC}(t_a; t_b)$ -values into z-scores, we consider a TFBS pair to be significant in the entire set of sequences, if the pair has a z-score  $\geq 3$ .

#### Additional files

**Additional file 1: PC-TraFF analysis with different distance constrains.** Significant pairs found by PC-TraFF using different distance constrains in the sequences of human RefSeq genes. (XLS 13 kb)

**Additional file 2: Synthetic sequences.** Synthetic sequences with USF-IRF1 binding sites. (FASTA 55 kb)

**Additional file 3: Genome-wide analysis in the context of RefSeq genes.** PC-TraFF significant pairs identified in the genome-wide promoter sequences of RefSeq genes. (XLS 4 kb)

**Additional file 4: Cell lines and tissues.** Cell lines and tissues, which are used to predict promoter sequences of miRNAs. (XLS 4 kb)

**Additional file 5: Genome-wide analysis in the context of miRNA genes.** PC-TraFF significant pairs identified in the genome-wide promoter sequences of miRNA genes. (XLS 3 kb)

**Additional file 6: Breast cancer analysis in the context of RefSeq genes.** PC-TraFF significant pairs identified in the promoter sequences of breast cancer-associated RefSeq genes. (XLS 12 kb)

**Additional file 7: Breast cancer analysis in the context of miRNA genes.** PC-TraFF significant pairs identified in the promoter sequences of breast cancer-associated miRNA genes. (XLS 13 kb)

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

MG developed the model underlying PC-TraFF. EW and SW adjusted the model together with MG. CM developed the model together with MG, designed and implemented the tool and interpreted the results together with RT, MG, and EW. XH was involved in the determination of TSSs for miRNAs. RT studied the interactions between TFs in promoter regions of miRNAs genes and interpreted the results together with CM, MG, and EW. MG conceived of and managed the project and wrote the final version of the manuscript. All authors read and approved the final version.

#### Acknowledgments

We thank our colleagues Martin Haubrock, Dariusz Wlochowitz, and Sebastian Zeidler for their helpful advice and insights at early stages of this project and for comments on the results.

**Author details**

<sup>1</sup>Institute of Bioinformatics, University of Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany. <sup>2</sup>Institute of Computer Science, University of Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany.

Received: 24 March 2015 Accepted: 17 November 2015

Published online: 01 December 2015

**References**

- Amoutzias GD, Robertson DL, de Peer YV, Oliver SG. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci.* 2008;33(5):220–9. <http://www.sciencedirect.com/science/article/pii/S0968000408000625>.
- Zhu Z, Shendure J, Church GM. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.* 2005;15(6):848–55. <http://genome.cshlp.org/content/15/6/848.abstract>.
- Mysickova A, Vingron M. Detection of interacting transcription factors in human tissues using predicted DNA binding affinity. *BMC Genomics.* 2012;13(Suppl 1):S2. <http://www.biomedcentral.com/1471-2164/13/S1/S2>.
- Navarro C, Lopez FJ, Cano C, Garcia-Alcalde F, Blanco A. CisMiner: Genome-wide in-silico cis-regulatory module prediction by fuzzy itemset mining. *PLoS ONE.* 2014;9(9):e108065. <http://dx.doi.org/10.1371/journal.pone.0108065>.
- Jankowski A, Prabhakar S, Tiurny J. TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics.* 2014;15:208. <http://www.biomedcentral.com/1471-2164/15/208>.
- Deyneko I, Kel A, Kel-Margoulis O, Deineko E, Wingender E, Weiss S. MatrixCatch - a novel tool for the recognition of composite regulatory elements in promoters. *BMC Bioinformatics.* 2013;14:241. <http://www.biomedcentral.com/1471-2105/14/241>.
- Nandi S, Blais A, Ioshikhes I. Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Res.* 2013;41(19):8822–41. <http://nar.oxfordjournals.org/content/41/19/8822.abstract>.
- Ha N, Polychronidou M, Lohmann I. COPS: Detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. *PLoS ONE.* 2012;7(12):e52055. <http://dx.doi.org/10.1371/journal.pone.0052055>.
- Sun H, Guns T, Fierro AC, Thorrez L, Nijssen S, Marchal K. Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection. *Nucleic Acids Res.* 2012;40(12):e90. <http://nar.oxfordjournals.org/content/40/12/e90.abstract>.
- Sun H, De Bie T, Storms V, Fu Q, Dholander T, Lemmens K, et al. ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules. *BMC Bioinformatics.* 2009;10(Suppl 1):S30. <http://www.biomedcentral.com/1471-2105/10/S1/S30>.
- Hu Z, Hu B, Collins J. Prediction of synergistic transcription factors by function conservation. *Genome Biol.* 2007;8(12):R257. <http://genomebiology.com/2007/8/12/R257>.
- Frith MC, Li MC, Weng Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* 2003;31(13):3666–8. <http://nar.oxfordjournals.org/content/31/13/3666.abstract>.
- Sinha S, van Nimwegen E, Siggia ED. A probabilistic method to detect regulatory modules. *Bioinformatics.* 2003;19(suppl 1):i292–301. [http://bioinformatics.oxfordjournals.org/content/19/suppl\\_1/i292.abstract](http://bioinformatics.oxfordjournals.org/content/19/suppl_1/i292.abstract).
- Frith MC, Hansen U, Weng Z. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics.* 2001;17(10):878–89. <http://bioinformatics.oxfordjournals.org/content/17/10/878.abstract>.
- Van Loo P, Marynen P. Computational methods for the detection of cis-regulatory modules. *Briefings in Bioinformatics.* 2009;10(5):509–24. <http://bib.oxfordjournals.org/content/10/5/509.abstract>.
- Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet.* 2012;13(7):469–83. <http://dx.doi.org/10.1038/nrg3242>.
- Hu Z, Gallo S. Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics.* 2010;11:49. <http://www.biomedcentral.com/1471-2164/11/49>.
- Pickert L, Reuter I, Klawonn F, Wingender E. Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics.* 1998;14(3):244–51. <http://bioinformatics.oxfordjournals.org/content/14/3/244.abstract>.
- Kel-Margoulis O, Kel A, Reuter I, Deineko I, Wingender E. TRANSCOMP: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* 2002;30:332–4.
- S A, Kaimal R. Document summarization using positive pointwise mutual information. *CoRR, Intl J Comput Sci Inf Technol (IJCSIT).* 2012;4, abs/1205.1638(2): <http://arxiv.org/abs/1205.1638>.
- Bouma G. Normalized (Pointwise) Mutual Information in Collocation Extraction. In: *Proceedings of the Biennial Conference of GSCL; 2009.* p. 31–40.
- Islam A, Inkpen D. Second order co-occurrence PMI for determining the semantic similarity of words. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, Italy; 2006. p. 1033–8.
- Damani OP. Improving Pointwise Mutual Information (PMI) by Incorporating Significant Co-occurrence. *CoRR.* In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* Vol. abs/1307.0596. Washington: Seattle. p. 163–169. <http://arxiv.org/abs/1307.0596>. Accessed 2 Jul 2013.
- Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics.* 2008;24(3):333–40.
- Kel A, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis O, Wingender E. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 2003;31(13):3576–9.
- Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.* 2008;9(4):326–32.
- Girgis H, Ovcharenko I. Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. *BMC Bioinformatics.* 2012;13:25. <http://www.biomedcentral.com/1471-2105/13/25>.
- Joshi H, Nord S, Frigessi A, Borresen-Dale AL, Kristensen V. Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes. *BMC Genomics.* 2012;13:199. <http://www.biomedcentral.com/1471-2164/13/199>.
- Chatr-aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 2014. URL <http://nar.oxfordjournals.org/content/early/2014/11/26/nar.gku1204.abstract>.
- Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43(D1):D447–52. <http://nar.oxfordjournals.org/content/43/D1/D447.abstract>.
- Yu X, Lin J, Zack DJ, Qian J. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* 2006;34(17):4925–36. <http://nar.oxfordjournals.org/content/34/17/4925.abstract>.
- Wingender E, Schoeps T, Dönitz J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* 2013;41(D1):D165–D170. <http://nar.oxfordjournals.org/content/41/D1/D165.abstract>.
- Kaczynski J, Cook T, Urrutia R. Sp1- and Krüppel-like transcription factors. *Genome Biol.* 2003;4(2):206. <http://genomebiology.com/2003/4/2/206>.
- Beishline K, Azizkhan-Clifford J. Sp1 and the "Hallmarks of Cancer". *FEBS J.* 2014. n/a–n/a, URL <http://dx.doi.org/10.1111/febs.13148>.
- Song CZ, Keller K, Murata K, Asano H, Stamatoyannopoulos G. Functional interaction between coactivators CBP/p300, PCAF, and transcription factor KLF2. *J Biol Chem.* 2002;277:7029–36. [This study shows the interaction of KLF13 with coactivators].
- Zhang W, Kadam S, Emerson B, Bieker J. Site-specific acetylation by p300 or CREB binding protein regulates erythroid Krüppel-like factor transcriptional activity via its interaction with the SWI-SNF complex. *Mol Cell Biol.* 2001;21:2413–22. [These results demonstrate that the acetylation of EKLF by p300/CBP is critical for optimal KLF1 activity].
- Whitfield T, Wang J, Collins P, Partridge EC, Aldred S, Trinklein N, et al. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 2012;13(9):R50. <http://genomebiology.com/2012/13/9/R50>.

38. Darnell JE. STATs and gene regulation. *Science*. 1997;277(5332):1630–5. <http://www.sciencemag.org/content/277/5332/1630.abstract>.
39. Levy D, Darnell J. Stats: transcriptional control and biological impact. *Nat Rev Mol Cell Biol*. 2002;3:651–62.
40. Goenka S, Kaplan M. Transcriptional regulation by STAT6. *Immunol Res*. 2011;50:87–96. <http://dx.doi.org/10.1007/s12026-011-8205-2>.
41. Dittmer J. The Biology of the Ets1 Proto-Oncogene. *Mol Cancer*. 2003;2:29. <http://www.molecular-cancer.com/content/2/1/29>.
42. Findlay VJ, LaRue AC, Turner DP, Watson PM, Watson DK. Understanding the role of ETS-mediated gene regulation in complex biological processes. 2013;119:1–61. <http://www.sciencedirect.com/science/article/pii/B9780124071902000010>.
43. Obika S, Reddy SY, Bruice TC. Sequence specific DNA Binding of Ets-1 transcription factor: molecular dynamics study on the Ets domain-DNA complexes. *J Mol Biol*. 2003;331(2):345–59. <http://www.sciencedirect.com/science/article/pii/S0022283603007265>.
44. Baillat D, Bègue A, Stéhelin D, Aumercier M. ETS-1 Transcription Factor Binds Cooperatively to the Palindromic Head to Head ETS-binding Sites of the Stromelysin-1 Promoter by Counteracting Autoinhibition. *J Biol Chem*. 2002;277(33):29386–98. <http://www.jbc.org/content/277/33/29386.abstract>.
45. Nakazawa Y, Suzuki M, Manabe N, Yamada T, Kihara-Negishi F, Sakurai T, et al. Cooperative interaction between ETS1 and GFI1 transcription factors in the repression of Bax gene expression. *Oncogene*. 2007;26(24):3541–50.
46. Karin M, gang Liu Z, Zandi E. AP-1 function and regulation. *Curr Opin Cell Biol*. 1997;9(2):240–46. <http://www.sciencedirect.com/science/article/pii/S0955067497800683>.
47. Han B, Rorke EA, Adhikary G, Chew YC, Xu W, Eckert RL. Suppression of AP1 transcription factor function in Keratinocyte suppresses differentiation. *PLoS ONE*. 2012;7(5):e36941. <http://dx.doi.org/10.1371>.
48. Hess J, Angel P, Schorpp-Kistner M. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci*. 2004;117(25):5965–73. <http://jcs.biologists.org/content/117/25/5965.abstract>.
49. Chinenov Y, Kerppola TK. Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene*. 2001;20(19):2438–52.
50. Ramirez-Carrozzi VR, Kerppola TK. Control of the orientation of Fos-Jun binding and the transcriptional cooperativity of Fos-Jun-NFAT1 complexes. *J Biol Chem*. 2001;276(24):21797–808. <http://www.jbc.org/content/276/24/21797.abstract>.
51. Block K, Shou Y, Poncz M. An Ets/Sp1 interaction in the 5'-flanking region of the megakaryocyte-specific alpha IIb gene appears to stabilize Sp1 binding and is essential for expression of this TATA-less gene. *Blood*. 1996;88(6):2071–80.
52. Sahoo A, Lee CG, Jash A, Son JS, Kim G, Kwon HK, et al. Stat6 and c-Jun Mediate Th2 Cell-Specific IL-24 Gene Expression. *J Immunol*. 2011;186(7):4098–109. <http://www.jimmunol.org/content/186/7/4098.abstract>.
53. Yin Q, Wang X, McBride J, Fewell C, Flemington E. B-cell Receptor Activation Induces BIC/miR-155 Expression through a Conserved AP-1 Element. *J Biol Chem*. 2008;283(5):2654–62. <http://www.jbc.org/content/283/5/2654.abstract>.
54. Harris TA, Yamakuchi M, Kondo M, Oettgen P, Lowenstein CJ. Ets-1 and Ets-2 regulate the expression of microRNA-126 in endothelial cells. *Arterioscler Thromb Vasc Biol*. 2010;30(10):1990–7.
55. Dubey R, Saini N. STAT6 silencing up-regulates cholesterol synthesis via miR-197/FOXJ2 axis and induces ER stress-mediated apoptosis in lung cancer cells. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. 2015;1849:32–43. <http://www.sciencedirect.com/science/article/pii/S1874939914002600>.
56. Liu D, Tao T, Xu B, Chen S, Liu C, Zhang L, et al. MiR-361-5p acts as a tumor suppressor in prostate cancer by targeting signal transducer and activator of transcription-6(STAT6). *Biochem Biophys Res Commun*. 2014;445:151–6. <http://www.sciencedirect.com/science/article/pii/S0006291X14001752>.
57. Bindra RS, Gibson SL, Meng A, Westermark U, Jasin M, Pierce AJ, et al. Hypoxia-Induced Down-regulation of BRCA1 Expression by E2Fs. *Cancer Res*. 2005;65(24):11597–604. <http://cancerres.aacrjournals.org/content/65/24/11597.abstract>.
58. Hynes N, Stoelzle T. Key signalling nodes in mammary gland development and cancer. *Myc. Breast Cancer Res*. 2009;11(5):210. <http://breast-cancer-research.com/content/11/5/210>.
59. Zhang W, Grivnickov S. Top Notch cancer stem cells by paracrine NF-kappaB signaling in breast cancer. *Breast Cancer Res*. 2013;15(5):316. <http://breast-cancer-research.com/content/15/5/316>.
60. Doetzlhofer A, Rotheneder H, Lagger G, Koranda M, Kurtev V, Brosch G, et al. Histone Deacetylase 1 Can Repress Transcription by Binding to Sp1. *Mol Cell Biol*. 1999;19(8):5504–11. <http://mcb.asm.org/content/19/8/5504.abstract>.
61. Alvira CM. Nuclear factor-kappa-B signaling in lung development and disease: One pathway, numerous functions. *Birth Defects Res Part A: Clinical Mol Teratology*. 2014;100(3):202–16. <http://dx.doi.org/10.1002/bdra.23233>.
62. Switzer C, Cheng R, Ridnour L, Glynn S, Amb S, Wink D. Ets-1 is a transcriptional mediator of oncogenic nitric oxide signaling in estrogen receptor-negative breast cancer. *Breast Cancer Res*. 2012;14(5):R125. <http://breast-cancer-research.com/content/14/5/R125>. [See related commentary by Marshall and Foster, <http://breast-cancer-research.com/content/14/6/113>].
63. Takai N, Miyazaki T, Nishida M, Shang S, Nasu K, Miyakawa I. Clinical relevance of Elf-1 overexpression in endometrial carcinoma. *Gynecol Oncol*. 2003;89(3):408–13. <http://www.sciencedirect.com/science/article/pii/S0090825803001318>.
64. Walker L, Fredericksen Z, Wang X, Tarrell R, Pankratz V, Lindor N, et al. Evidence for SMAD3 as a modifier of breast cancer risk in BRCA2 mutation carriers. *Breast Cancer Res*. 2010;12(6):R102. <http://breast-cancer-research.com/content/12/6/R102>.
65. Kim DW, Gazourian L, Quadri SA, Romieu-Mourez R, Sherr DH, Sonenshein GE. The RelA NF-kappaB subunit and the aryl hydrocarbon receptor (AhR) cooperate to transactivate the c-myc promoter in mammary cells. *Oncogene*. 2000;19(48):5498–5506.
66. Song J, Clagett-Dame M, Peterson RE, Hahn ME, Westler WM, Sicinski RR, et al. A ligand for the aryl hydrocarbon receptor isolated from lung. *Proc Natl Acad Sci*. 2002;99(23):14694–9. <http://www.pnas.org/content/99/23/14694.abstract>.
67. Champion CG, Labrie M, Grosset AA, St-Pierre Y. The CCAAT/enhancer-binding protein beta-2 isoform (CEBPβ-2) upregulates galectin-7 expression in human breast cancer cells. *PLoS ONE*. 2014;9(5):e95087.
68. Shah SN, Cope L, Poh W, Belton A, Roy S, Talbot CC, et al. HMGA1: a master regulator of tumor progression in triple-negative breast cancer cells. *PLoS ONE*. 2013;8(5):e63419.
69. Foti D, Iuliano R, Chiefari E, Brunetti A. A nucleoprotein complex containing Sp1, C/EBPβ, and HMGI-Y controls human insulin receptor gene transcription. *Mol Cell Biol*. 2003;23(8):2720–32. <http://mcb.asm.org/content/23/8/2720.abstract>.
70. George OL, Ness SA. Situational awareness: regulation of the Myb transcription factor in differentiation, the cell cycle and oncogenesis. *Cancers*. 2014;6(4):2049–71. <http://www.mdpi.com/2072-6694/6/4/2049>.
71. Shen Q, Uray IP, Li Y, Krisko TI, Strecker TE, Kim HT, et al. The AP-1 transcription factor regulates breast cancer cell growth via cyclins and E2F factors. *Oncogene*. 2008;27(3):366–77.
72. Wei M, Liu B, Gu Q, Su L, Yu Y, Zhu Z. Stat6 cooperates with Sp1 in controlling breast cancer cell proliferation by modulating the expression of p21(Cip1/WAF1) and p27 (Kip1). *Cell Oncol (Dordr)*. 2013;36:79–93.
73. Gooch JL, Christy B, Yee D. STAT6 mediates interleukin-4 growth inhibition in human breast cancer cells. *Neoplasia*. 2002;4(4):324–31.
74. Foxler DE, James V, Shelton SJ, Vallim TQ, Shaw PE, Sharp TV. PU.1 is a major transcriptional activator of the tumour suppressor gene LIMS1. *FEBS Lett*. 2011;585(7):1089–96.
75. Mattia G, Errico MC, Felicetti F, Petrini M, Bottero L, Tomasello L, et al. Constitutive activation of the ETS-1-miR-222 circuitry in metastatic melanoma. *Pigment Cell Melanoma Res*. 2011;24(5):953–65.
76. Tavazoie SF, Alarcon C, Oskarsson T, Padua D, Wang Q, Bos PD, et al. Endogenous human microRNAs that suppress breast cancer metastasis. *Nature*. 2008;451(7175):147–52.
77. Tschan MP, Reddy VA, Ressa A, Arvidsson G, Fey MF, Torbett BE. PU.1 binding to the p53 family of tumor suppressors impairs their transcriptional activity. *Oncogene*. 2008;27(24):3489–3493.
78. Okuno Y, Yuki H. PU.1 is a tumor suppressor for B cell malignancies. *Oncotarget*. 2012;3(12):1495–6.
79. Sun Y, Sun J, Tomomi T, Nieves E, Mathewson N, Tamaki H, et al. PU.1-dependent transcriptional regulation of miR-142 contributes to its hematopoietic cell-specific expression and modulation of IL-6. *J Immunol*. 2013;190(8):4005–13.



80. Rosa A, Ballarino M, Sorrentino A, Sthandier O, De Angelis FG, Marchioni M, et al. The interplay between the master transcription factor PU.1 and miR-424 regulates human monocyte/macrophage differentiation. *Proc Natl Acad Sci U S A*. 2007;104(50):19849–54.
81. Fujita S, Ito T, Mizutani T, Minoguchi S, Yamamichi N, Sakurai K, et al. miR-21 Gene expression triggered by AP-1 is sustained through a double-negative feedback mechanism. *J Mol Biol*. 2008;378(3):492–504.
82. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. *Cancer Res*. 2005;65(16):7065–70.
83. Sato M, Morii E, Komori T, Kawahata H, Sugimoto M, Terai K, et al. Transcriptional regulation of osteopontin gene in vivo by PEBP2 $\alpha$ /CBFA1 and ETS1 in the skeletal tissues. *Oncogene*. 1998;17(12):1517–25.
84. He B, Mirza M, Weber GF. An osteopontin splice variant induces anchorage independence in human breast cancer cells. *Oncogene*. 2006;25(15):2192–202.
85. Mikita T, Kurama M, Schindler U. Synergistic Activation of the Germline  $\epsilon$  Promoter Mediated by Stat6 and C/EBP $\beta$ . *J Immunol*. 1998;161(4):1822–8.
86. Chand AL, Herridge KA, Thompson EW, Clyne CD. The orphan nuclear receptor LRH-1 promotes breast cancer motility and invasion. *Endocr Relat Cancer*. 2010;17(4):965–75. <http://erc.endocrinology-journals.org/content/17/4/965.abstract>.
87. Hwang-Verslues WW, Chang PH, Wei PC, Yang CY, Huang CK, Kuo WH, et al. miR-495 is upregulated by E12/E47 in breast cancer stem cells, and promotes oncogenesis and hypoxia resistance via downregulation of E-cadherin and REDD1. *Oncogene*. 2011;30(21):2463–74.
88. Slyper M, Shahar A, Bar-Ziv A, Granit RZ, Hamburger T, Maly B, et al. Control of breast cancer growth and initiation by the stem cell-associated transcription factor TCF3. *Cancer Res*. 2012;72(21):5613–24.
89. Zhao M, Sun J, Zhao Z. Synergetic regulatory networks mediated by oncogene-driven microRNAs and transcription factors in serous ovarian cancer. *Mol Biosyst*. 2013;9(12):3187–98.
90. Guo Z, Maki M, Ding R, Yang Y, Zhang B, Xiong L. Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci Rep*. 2014;4:5150.
91. Delfino KR, Rodriguez-Zas SL. ranscription Factor-MicroRNA-Target Gene Networks Associated with Ovarian Cancer Survival and Recurrence. *PLoS ONE*. 2013;8(3):e58608. <http://dx.doi.org/10.1371>.
92. Le T, Liu L, Liu B, Tsykin A, Goodall G, Satou K, et al. Inferring microRNA and transcription factor regulatory networks in heterogeneous data. *BMC Bioinformatics*. 2013;14:92. <http://www.biomedcentral.com/1471-2105/14/92>.
93. De Martino I, Visone R, Fedele M, Petrocca F, Palmieri D, Martinez Hoyos J, et al. Regulation of microRNA expression by HMGA1 proteins. *Oncogene*. 2009;28(11):1432–42.
94. Mansueto G, Forzati F, Ferraro A, Pallante P, Bianco M, Esposito F, et al. Identification of a New Pathway for Tumor Progression: MicroRNA-181b Up-Regulation and CBX7 Down-Regulation by HMGA1 Protein. *Genes Cancer*. 2010;1(3):210–24.
95. Klepper K, Sandve G, Abul O, Johansen J, Drablos F. Assessment of composite motif discovery methods. *BMC Bioinformatics*. 2008;9:123. <http://www.biomedcentral.com/1471-2105/9/123>.
96. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32(suppl 1):D493–6.
97. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*. 2007;130:77–88.
98. Marsico A, Huska MR, Lasserre J, Hu H, Vucicevic D, Musahl A, et al. PROMiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol*. 2013;14(8):R84.
99. Hannehalli S, Levy S. Predicting transcription factor synergism. *Nucleic Acids Res*. 2002;30(19):4278–84. <http://nar.oxfordjournals.org/content/30/19/4278.abstract>.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



## **A.2. Removing background Co-occurrences of TFBSs greatly improves the prediction of specific TF cooperations**



# Removing Background Co-occurrences of Transcription Factor Binding Sites Greatly Improves the Prediction of Specific Transcription Factor Cooperations

Cornelia Meckbach<sup>1\*</sup>, Edgar Wingender<sup>1</sup> and Mehmet Gültas<sup>1,2,3</sup>

<sup>1</sup> Institute of Bioinformatics, University Medical Center Göttingen, Georg-August-University Göttingen, Göttingen, Germany,

<sup>2</sup> Department of Breeding Informatics, Georg-August University Göttingen, Göttingen, Germany, <sup>3</sup> Center for Integrated Breeding Research (CiBreed), Georg-August University Göttingen, Göttingen, Germany

## OPEN ACCESS

### Edited by:

Alexandre V. Morozov,  
Rutgers University, The State  
University of New Jersey,  
United States

### Reviewed by:

Vladimir B. Teif,  
University of Essex, United Kingdom  
Hauke Busch,  
Universität zu Lübeck, Germany

### \*Correspondence:

Cornelia Meckbach  
cornelia.meckbach@  
bioinf.med.uni-goettingen.de

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 16 February 2016

**Accepted:** 08 May 2018

**Published:** 29 May 2018

### Citation:

Meckbach C, Wingender E and  
Gültas M (2018) Removing  
Background Co-occurrences of  
Transcription Factor Binding Sites  
Greatly Improves the Prediction of  
Specific Transcription Factor  
Cooperations. *Front. Genet.* 9:189.  
doi: 10.3389/fgene.2018.00189

Today, it is well-known that in eukaryotic cells the complex interplay of transcription factors (TFs) bound to the DNA of promoters and enhancers is the basis for precise and specific control of transcription. Computational methods have been developed for the identification of potentially cooperating TFs through the co-occurrence of their binding sites (TFBSs). One challenge of these methods is the differentiation of TFBS pairs that are specific for a given sequence set from those that are ubiquitously appearing, rendering the results highly dependent on the choice of a proper background set. Here, we present an extension of our previous PC-TraFF approach that estimates the background co-occurrence of any TF pair by preserving the (oligo-) nucleotide composition and, thus, the core of TFBSs in the sequences of interest. Applying our approach to a simulated data set with implanted TFBS pairs, we could successfully identify them as sequence-set specific under a variety of conditions. When we analyzed the gene expression data sets of five breast cancer associated subtypes, the number of overlapping pairs could be dramatically reduced in comparison to our previous approach. As a result, we could identify potentially cooperating transcriptional regulators that are characteristic for each of the five breast cancer subtypes. This indicates that our approach is able to discriminate specific potential TF cooperations against ubiquitously occurring combinations. The results obtained with our method may help to understand the genetic programs governing specific biological processes such as the development of different tumor types.

**Keywords:** transcription factor (TF), TF cooperations, sequence-set specific TF cooperations, background correction, TF co-occurrences

## 1. INTRODUCTION

Transcription factors (TFs) are a special class of cellular proteins that are essential for controlling different genetic programs such as adaption to the environment, immune response, organogenesis or embryonic development by regulating gene expression. The human genome encodes roughly 1500–2000 different TFs which bind to short degenerate DNA motifs, known as transcription factor

binding sites (TFBSs). In higher organisms, the binding of TFs occurs in a specific combination within DNA regulatory regions (promoters as well as distal elements, such as enhancers) to form purposive dimers or higher order complexes to activate or repress their target genes. Due to the fact that eukaryotic DNA is packed in chromatin, TFs show additionally competing or cooperative DNA binding with chromatin associated proteins (Teif and Rippe, 2010). Besides this, based on the co-occurrence of their TFBSs TFs exert functional cooperations which play an important role in the regulation of the different genetic programs in mammals (Boyer et al., 2005; Hu and Gallo, 2010; Neph et al., 2012). Today, it is well-known that the selection of cooperation partners for TFs depends on their biological functions, e.g., cell cycle control, cell homeostasis, or cell differentiation in different cell types. As a result of these properties, TFs change their partners to specify their functions according to the cellular context.

In the last decade, a various number of computational methods for the identification of cooperating TFs has been proposed (Hu et al., 2007; Van Loo and Marynen, 2009; Girgis and Ovcharenko, 2012; Ha et al., 2012; Sun et al., 2012; Deyneko et al., 2013; Nandi et al., 2013; Jankowski et al., 2014; Navarro et al., 2014; Meckbach et al., 2015; Wu and Lai, 2016; Spadafore et al., 2017). Among these methods, predicting the putative TFBSs in the sequences under study and building a meaningful quantification measure of the cooperation between two TFs are two essential steps to make the predictions successful. Based on these steps, different strategies/ideas have been used for the identification of cooperating TF pairs such as the TFBS co-occurrences of cooperative pairs are more often than expected by chance and have significantly closer distances. In this context, several methods such as statistical methods like the hypergeometric test, clustering approaches, randomized occurrence frequency model (OF<sub>r</sub>) or Markov models have been developed (Hu et al., 2007; Chuang et al., 2009; Girgis and Ovcharenko, 2012; Ha et al., 2012; Mysickova and Vingron, 2012; Sun et al., 2012; Nandi et al., 2013; Jankowski et al., 2014; Lai et al., 2014; Navarro et al., 2014; Spadafore et al., 2017).

Employing a comprehensive performance evaluation study on the prediction results of those methods, Lai et al. (2014) have shown that the success rates of different approaches strongly depend on the corresponding evaluation criteria. This finding is also supported by our results, which we have presented in Meckbach et al. (2015). However, the predictions of almost all of these methods suffer from many types of obstacles that might occur as a result of high background like common regulatory programs between cell types and the environmental components in their regulatory sequences like GC content or nucleotide composition - indicating the ratio of the constituent monomer units/bases- as well as the noise effect of false positive putative TFBSs. Hence, such obstacles lead into background co-occurrence of TFBSs and consequently the results of a certain method are often highly overlapping for different sequence sets. Zeidler et al. (2016) have clearly demonstrated this problem in their study for detection of stage-specific TF pairs in a time series data set during heart development. To overcome this problem, they have further applied Markov clustering algorithm

(MCL) (Dongen, 2000) to the pairs predicted by MatrixCatch methodology (Deyneko et al., 2013). Although several negligible TF cooperations could be eliminated, the application of MCL algorithm in this context is only based on the observed frequencies of TFBSs and does not consider the sequence specific environmental components. Consequently, the results of this approach seem to be conservative and not sequence set specific, yet.

To deal with this problem to some extent, we applied in our previous study the average product correction (APC) theorem (Dunn et al., 2008) in order to determine for each TFBS pair their background co-occurrence resulting from their possibly false positive TFBS predictions in the entire sequence set under study. Although, with respect to APC theorem, the background noise effect of false positive TFBSs could be successfully eliminated in the detection of significant TF pairs, the power and functionality of APC theorem appears to be insufficient to handle the remaining obstacles for the identification of sequence-set specific TF cooperations. In order to overcome the missing point of PC-TraFF workflow (Meckbach et al., 2015), we propose in this study an efficient approach that accurately quantifies the level of background co-occurrence of two TFBSs considering different types of obstacles (mentioned above) in the sequences under study. For this purpose, by preserving the (oligo-) nucleotide composition of the sequences of interest, we create a sufficient number of new shuffled sequence sets and based on these sets the background co-occurrence of a TFBS pair is measured. This process ensures that TF cooperations, which are very sensitive regarding the context of nucleotides and the distance of their binding sites, will become remarkable small background-values in comparison to common (ubiquitously occurring) TF pairs. These ubiquitously occurring TF pairs are often found as significant for different sequence sets and are less susceptible to the behavior of their binding sites in the set of sequences. Consequently, removal of this background leads to the separation of sequence set-specific TF pairs from the common ones.

To demonstrate the performance and functionality of our proposed approach, we analyzed a simulation data set as well as five breast cancer subtype-associated gene sets, and present the results step by step by providing comparative analysis. These data sets have been chosen because the importance of cooperating TF pairs have been well-studied in Meckbach et al. (2015).

## Terminology

For the sake of simplicity, we adapt the terminology of our previous paper (Meckbach et al., 2015). In doing so, each match of a position weight matrix (PWM) with a segment of genomic DNA is called a (potential) *transcription factor binding site* (TFBS). TFBSs are represented by names of their corresponding PWMs. The PWMs of TRANSFAC (Wingender, 2008) used in this report are denoted with their TRANSFAC identifiers, the structure of which is:  $V\$factormname\_version$ , where “V\$” indicates that the PWM is representing a TFBS of a vertebrate TF. *factormname* refers to the TF name, while there are more than one PWM representing the binding motif of a certain factor, *version* is required for the unambiguous identification of the PWM. TFBS pairs refer to co-occurring TFBSs. It is important to note that we

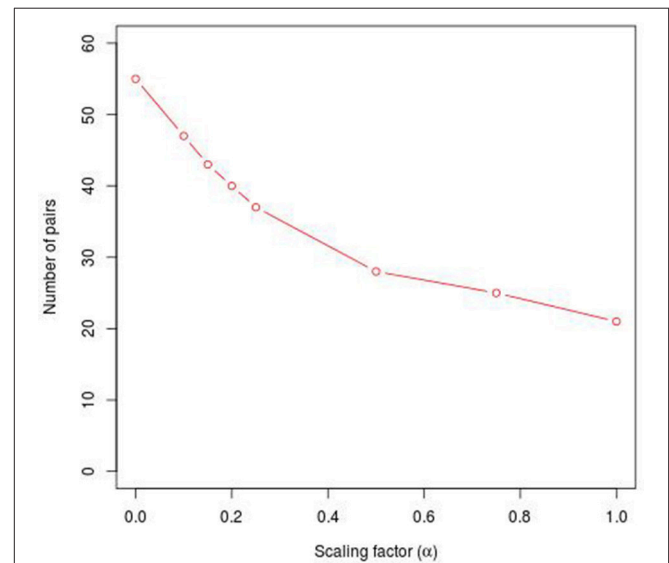
cannot make any statement about the kind of interaction such co-occurrence may be associated with (cooperativity, synergistic or antagonistic interaction etc.). The term cooperation refers to any kind of functional cooperation and/or physical interaction between the constituents of the predicted TFBS pairs.

## 2. RESULTS AND DISCUSSION

In this study, we introduce an extension of our previous methodological approach PC-TraFF for the separation of sequence-set specific cooperating transcription factors based on the co-occurrence of their binding sites from common ones. The overall workflow of our approach comprises two parts. First, the original PC-TraFF algorithm is used in order to predict significant TFBS pairs in a set of sequence where PC-TraFF provides for each significant TFBS pair  $t_a$  and  $t_b$  a pointwise mutual information score  $\text{PMI}_{pc}^{APC}(t_a; t_b)$ . Thereby, the minimal and maximal distance threshold for two TFBSs to form a pair is set to 5 and 20 bp, respectively, in order to provide a proper comparison to the original PC-TraFF-results.

Second, in order to separate PC-TraFF significant TFBS pairs into the two groups of sequence-set specific and common (generally important) combinations, we apply our extension approach. For this purpose, out of the sequences of interest, a sufficiently large number of background sets is created by shuffling the original sequences, whereby the general nucleotide composition of the sequences as well as the core of the putative TFBSs are maintained. For all these background sets, the original PC-TraFF algorithm is applied to calculate  $\text{PMI}_{pc}^{APC}$ -values between all TFBS pairs. Afterwards, using these values the level of average background cooperation, which is defined as  $\text{AVG}(\text{PMI}(t_a; t_b))$ -value, between two TFs based on their binding sites over all sets of background sequences is calculated. The subtraction of  $\text{AVG}(\text{PMI})$ -values from their initial  $\text{PMI}_{pc}^{APC}$ -values results in the separation of sequence-set specific pairs from the common co-occurrences. To this end, we additionally introduced a factor  $\alpha \in [-1, 1]$  to enlarge/reduce the effect of the subtracted background level by linearly influencing the subtracted average value  $\text{AVG}(\text{PMI}(t_a; t_b))$ . If  $\alpha = 1$ , the  $2 \times \text{AVG}(\text{PMI}(t_a; t_b))$ -value is subtracted from the initial  $\text{PMI}_{pc}^{APC}$ -value,  $\alpha = 0$  results simply in the subtraction of the observed  $\text{AVG}(\text{PMI}(t_a; t_b))$  value, while an  $\alpha$ -value of  $-1$  results in the original PC-TraFF predictions. Thus,  $\alpha$  enlarges/reduces the level of the subtracted background and is thereby influencing the number of identified specific pairs. However, our results suggest that the impact of  $\alpha$  on the number of specific pairs strongly depends on the individual sequence sets and appears not to be linear (e.g., see **Figure 1**) although the factor itself has a linear influence on the subtracted background level.

It is important to note that the Results section of this study mainly considers the influence of our proposed extension approach on the cooperating TFs identified by the PC-TraFF algorithm. Researchers, who are interested in the biological functions of individual TF cooperations, are kindly referred to the original PC-TraFF paper (Meckbach et al., 2015).



**FIGURE 1** | Number of specific TFBS pairs for the synthetic sequence set in dependence on different  $\alpha$ -values. The synthetic sequence set consists of 200 sequences of length 1000 bps, each of these sequences contains artificially inserted binding site pairs (V\$IRF1\_01 - V\$USF1\_01) for the cooperation between transcription factors IRF1 and USF1 with a minimal distance of 5 bp and a maximal distance of 20 bp. The  $\alpha$ -value linearly influences the subtracted background level (e.g.,  $\alpha = 0$  results in the subtraction of the  $\text{AVG}(\text{PMI}(t_a; t_b))$  value,  $\alpha = 1$  indicates the subtraction of the  $2 \times \text{AVG}(\text{PMI}(t_a; t_b))$ -value).

**TABLE 1** | Total number of specific TFBS pairs for the simulation data set using different  $\alpha$ -values.

$\alpha$ -value	Rank of artificially inserted pair	Total number of pairs found
$\alpha = -1$	18	58
$\alpha = 0$	16	55
$\alpha = 0.1$	15	47
$\alpha = 0.15$	14	43
$\alpha = 0.2$	12	40
$\alpha = 0.25$	11	37
$\alpha = 0.5$	6	28
$\alpha = 0.75$	6	25
$\alpha = 1$	5	21

The rank according to z-score indicates the position of the inserted pair. The scaling factor  $\alpha = -1$  indicates the significant TFBS pairs identified by the original PC-TraFF algorithm.

### 2.1. Analysis of Simulation Data

Analyzing the sequences in the simulation data set, the original PC-TraFF algorithm identified 58 TFBS pairs as significant ( $\alpha = -1$ ), where the artificially inserted binding site pair of the cooperating transcription factors IRF1 and USF1 is on position 18 according to z-score ranking. However, applying our extension approach to the results of PC-TraFF, only three of the 58 significant pairs were determined as common ones (see **Table 1**) based on the calculated background co-occurrence of TFBSs ( $\alpha = 0$ ). This rather low number of common pairs indicates that in a unspecific sequence set, the quantification

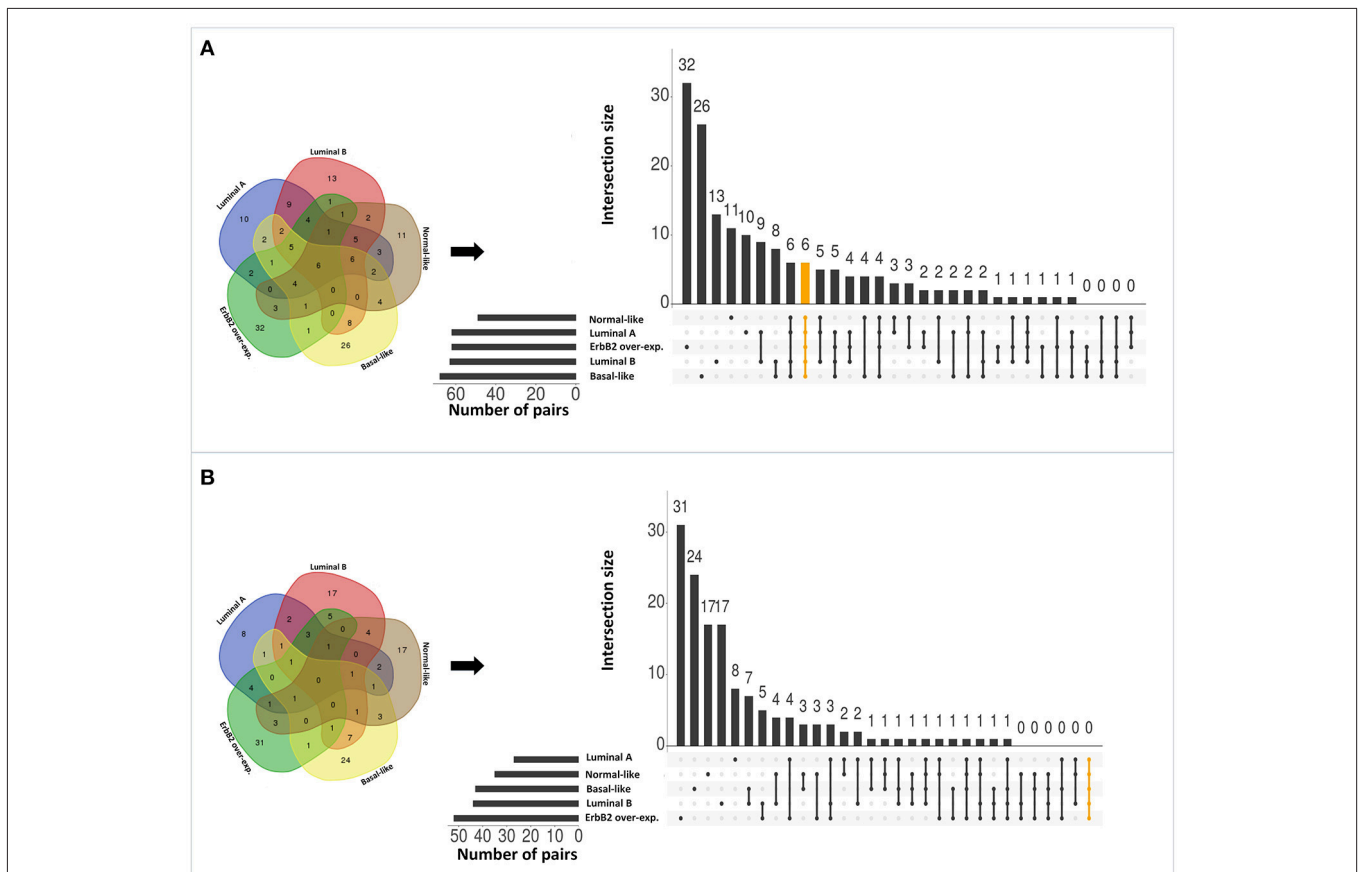
of correct background could be difficult which, in the worst case, may cause that sequence-set specific cooperations cannot be separated from common ones. To overcome this problem, the consideration of the scaling factor  $\alpha$  is important. **Figure 1** shows the influence of  $\alpha$  on the results. Although a variety of pairs are eliminated by means of different scaling factors, the inserted pair has been identified as sequence-set specific for each  $\alpha$ -value. Considering the  $z$ -score ranking of TFBS pairs, the position of the inserted pair is rising with an increasing  $\alpha$ -value (see **Table 1**). It has to be noted that the inserted binding sites are also matched by other PWMs, resulting in a variety of additional artificially arising TFBS pairs that consequently appear to be specific for the given sequence set.

### 2.2. Analysis of Breast Cancer Subtype Associated Promoter Sequences

Applying the original PC-TraFF algorithm to each BRC-subtype associated promoter sequences, we observed: (i) 62 TFBS pairs for *Luminal A*; (ii) 63 pairs for *Luminal B*; (iii) 68 pairs for *Basal-like*; (iv) 49 pairs for *Normal-like*; and (v) 62 pairs for *ErbB2 over-expressing* data set as significant. A comparison between these pairs shows that there are several pairs found as significant

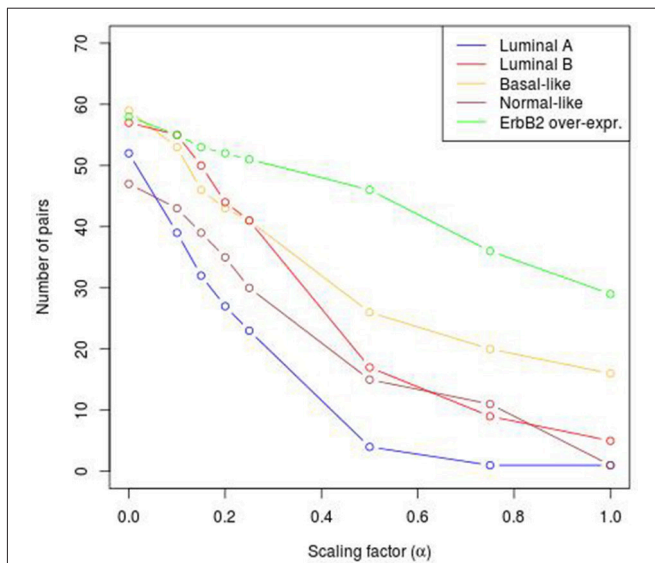
for more than one BRC-subtype (see **Figure 2A**), although the promoter sequences in all subtypes are unique (not overlapping). The reason of these overlapping pairs could be due to the same origin of the data and common regulatory programs which interfere with the identification of BRC-subtype specific TF cooperations.

To reveal the BRC-subtype specific TF cooperations, we additionally applied our extension approach using different  $\alpha$ -values to these significant pairs. The results of this analysis indicate that the scaling factor  $\alpha$  dramatically influences the number of sequence-set specific TFBS pairs. For example, on average 90% of the significant pairs have been determined as sequence-set specific by setting  $\alpha = 0$ , and 66% or 35% of significant pairs are assigned as sequence-set specific by setting  $\alpha = 0.2$  or  $\alpha = 0.5$ , respectively (**Figure 3**). Further, **Figure 3** shows that, the influence of the scaling factor  $\alpha$  is not consistent between the different sequence sets. While the number of specific TFBS pairs detected for *Luminal A* promoter sequences is dramatically decreasing and finally, 1% of all significant pairs have been determined as specific, the number of specific pairs for *ErbB2 over-expressing* promoter sequences has only slightly decreased in accordance with the increment of  $\alpha$ -value and in



**FIGURE 2 |** Number of significant TFBS pairs of five BRC-subtypes and their overlap represented in Venn diagrams and in matrix layouts using UpSet technique (Conway et al., 2017). Dark circles in the matrix layout indicate subtypes that are part on the intersection. Orange lines highlight the intersection between all BRC-subtypes. **(A)** Pairs identified by the original PC-TraFF version. **(B)** Sequence-set specific pairs determined by our extension approach using a scaling factor  $\alpha = 0.2$ .





**FIGURE 3 |** Number of sequence-set specific pairs found in the promoter sequences of differentially expressed genes of five BRC-subtypes depending on the  $\alpha$ -value. The  $\alpha$ -value linearly influences the subtracted background level (e.g.,  $\alpha = 0$  results in the subtraction of the mean,  $\alpha = 1$  indicates the subtraction of the  $2 \times \text{AVG}(\mathbb{P}^{\text{MII}}(t_a; t_b))$ -value).

an extreme case ( $\alpha = 1$ ) 47% of significant pairs in this subtype are assigned as specific. In addition, **Figure 2** depicts in detail for  $\alpha = 0.2$  the differences between significant and specific pairs for any BRC-subtype. By considering the sequence-set specific pairs, it is remarkable that like in the original PC-TraFF analysis, the *Luminal A* promoter sequence set has the lowest number of unique pairs (eight), and *ErbB2 over-expressing* promoter sequences have the largest number of unique TFBS pairs. The intersection of all BRC-subtypes specific pairs is zero.

Interestingly, after applying our extension approach, there are more sequence-set specific unique pairs for *Normal-like* and *Luminal B* subtypes (**Figure 2B**) than significant unique pairs (**Figure 2A**). For *Normal-like* data set, there are 11 significant and 17 specific unique pairs. In particular, six pairs that were identified in the original PC-TraFF analysis for several subtypes are determined to be solely sequence-set specific for *Normal-like* subtype. For example, the pairs (V\$CEBP\_Q2 - V\$HMGIIY\_Q6) and (V\$ELK1\_Q2 - V\$CETS1P54\_Q1) are significant for four different breast cancer subtypes or the pair (V\$CEBPB\_Q2 - V\$CEBP\_Q2) is significant in the original PC-TraFF version for three BRC-subtypes, but they are sequence-set specific only for *Normal-like* subtype (for details see **Table 2**).

For *Luminal B* subtype, 13 pairs were uniquely identified as significant by the original PC-TraFF algorithm and 17 pairs were uniquely assigned as specific. In this case, seven pairs that were common in the original PC-TraFF analysis have been determined to be sequence-set specific only for *Luminal B* subtype. Further, three of the unique significant pairs (V\$MYB\_Q5\_Q1 - V\$MAF\_Q6\_Q1, V\$NFKB\_Q6 - V\$CP2\_Q2, V\$HMGIIY\_Q6 - V\$MAF\_Q6\_Q1) were assigned as common co-occurrences according their negative  $\mathbb{P}^{\text{MII}}^{\text{specific}}$ -values.

**TABLE 2 |** Pairs that were identified as significant by PC-TraFF algorithm ( $\alpha = -1$ ) for different BRC-subtypes but are specific solely for a certain subtype using an  $\alpha$ -value of 0.2 for the background correction.

Specific for subtype	TFBS pairs	Significant in subtypes
Normal-like	V\$CEBPB_Q2 - V\$HMGIIY_Q6	<i>Basal-like, Luminal A, Luminal B, Normal-like</i>
	V\$ELK1_Q2 - V\$CETS1P54_Q1	<i>Basal-like, Luminal A, Luminal B, Normal-like</i>
	V\$CEBPB_Q2 - V\$CEBP_Q2	<i>ErbB2 over-expressing, Luminal B, Normal-like</i>
	V\$NFKB_Q6 - V\$SP1_Q4_Q1	<i>Luminal A, Normal-like</i>
	V\$EGR_Q6 - V\$AHRHIF_Q6	<i>Basal-like, Normal-like</i>
	V\$GR_Q6_Q1 - V\$PR_Q2	<i>ErbB2 over-expressing, Normal-like</i>
Luminal B	V\$CETS1P54_Q1 - V\$AHRHIF_Q6	<i>Luminal A, Luminal B, Normal-like</i>
	V\$E2F_Q3_Q1 - V\$PEBP_Q6	<i>Luminal A, Luminal B</i>
	V\$MYC_MAX_B - V\$AHRHIF_Q6	<i>Basal-like, Luminal A, Luminal B</i>
	V\$NFKB_Q6 - V\$E2F_Q3_Q1	<i>Luminal A, Luminal B</i>
	V\$NFKB_Q6 - V\$AHRHIF_Q6	<i>Luminal A, Luminal B</i>
	V\$CETS1P54_Q1 - V\$CP2_Q2	<i>Luminal A, Luminal B</i>
	V\$CETS1P54_Q1 - V\$MYC_MAX_B	<i>Basal-like, Luminal A, Luminal B, Normal-like</i>

Besides this, there are further six pairs identified by the original PC-TraFF algorithm as significant for all five BRC-subtypes, but they are assigned to be specific only for some of these subtypes (for details see **Figure 2** and **Table 3**). For example the TFBS pair (V\$CEBPB\_Q2 - V\$STAT6\_Q1) indicating the cooperation between the transcription factors CEBPB and STAT6 can still be found in the sequence-set specific pairs of *Luminal A*, *Luminal B* and *Basal-like* subtypes. In contrast, the pairs (V\$MYC\_MAX\_B - V\$E2F\_Q3\_Q1) and (V\$STAT6\_Q1 - V\$HMGIIY\_Q6) have been determined as specific only for *Basal-like* and *Normal-like* promoter sequence sets, respectively.

Finally, we built up cooperation networks based on the significant TFBS pairs, where the nodes refer to TFBSs and edges to predicted co-occurrences and thus, to cooperations between them, in order to demonstrate in an exemplary way the comparative analysis between the results of our extension approach and those of the original PC-TraFF algorithm. The cooperation network based on PC-TraFF significant TFBS pairs for *Luminal A* subtype (see **Figure 4**) consists of 33 nodes and 62 edges. Reducing the network by only considering sequence-set TFBS pairs results in the elimination of 7 nodes and 35 edges. Consequently, the remaining part of the network is built up of 26 nodes with their 27 sequence-set specific cooperations (edges). It is remarkable that some TFBSs that serve as hubs in the original network are still hub nodes in the reduced network but show a lower number of neighboring nodes (e.g., V\$CETS1P54\_Q1, V\$MYB\_Q5\_Q1, and V\$HMGIIY\_Q6). On the other side, there are some highly connected nodes of the original network that are missing in the specific pair network. For example the degree

**TABLE 3 |** TFBS pairs, which were identified as significant by original PC-TraFF algorithm for all five BRC-subtypes but were determined as specific only in certain subtypes.

TFBS pair	Specific for subtype(s)	Pairs documentation
V\$CETS1P54_01 - V\$ETS_Q4	<i>ErbB2 over-expressing, Luminal A</i>	BioGRID, TransCompel®
V\$MYC_MAX_B - V\$E2F_Q3_01	<i>Basal-like</i>	TransCompel®
V\$CEBPB_02 - V\$STAT6_01	<i>Luminal A, Luminal B, Basal-like</i>	TransCompel®
V\$STAT6_01 - V\$HMG1Y_Q6	<i>Normal-like</i>	-
V\$CETS1P54_01 - V\$NFKB_Q6	<i>Luminal A, Normal-like, Basal-like</i>	TransCompel®
V\$AP1_Q2_01 - V\$AP1_Q4_01	<i>Luminal A, Luminal B, ErbB2 over-expressing</i>	BioGRID, TransCompel®

The last column indicates the databases that document the evidence for these pairs. For this purpose, we used TRANSCompel® (Kel-Margoulis et al., 2002) and BioGRID interaction database (Chatr-aryamontri et al., 2014), which contain experimentally proven pairs.

of V\$NFKB\_Q6 or V\$AHRIF\_Q6 decreases from six neighbors to one neighbor and V\$SP1\_Q4\_01 is totally missing in the network of specific pairs. The node representing the binding site V\$SMAD\_Q6\_01 lost just one of its neighbors in this network and thereby, it is among the 25% nodes of highest degree.

A closer look at the cooperation network of significant TFBS pairs identified for the *Basal-like* data set discloses that 43 out of 68 significant pairs have been assigned to be sequence-set specific based on our extension approach with a scaling factor  $\alpha = 0.2$  (see **Figure 5A**). Setting  $\alpha = 0.5$  for this analysis leads to elimination of the vast majority of the pairs and consequently 16 pairs have been determined to be specific in the promoter sequences of *Basal-like* subtype (see **Figure 5B**). A comparison between cooperation networks of *Luminal A* and *Basal-like* subtypes suggests that by considering the same scaling factor our extension approach has more influence on significant pairs found for *Luminal A* data set than those found for *Basal-like* data set. The reason for this finding might be that *Basal-like* data set is more specific than *Luminal A* data set regarding to transcriptional regulation. Thus, the level of background co-occurrence of TFBSs resulting from common regulatory programs seems to be remarkable higher in *Luminal A* data set than those of *Basal-like* data set.

### 3. METHODS

#### 3.1. Data Sets

In order to assess the effectiveness of our approach and to present a detailed comparison with the results of original PC-TraFF algorithm, we analyzed in this study the data sets that have already been reported in Meckbach et al. (2015). The first data set is a simulation data set consisting of 200 sequences with the length of 1000 bps. Each of these sequences contains artificially inserted binding site pairs (V\$IRF1\_01 - V\$USF\_01) for the cooperation between transcription factors IRF1 and USF1 with a minimal distance of 5 bp and a maximal distance of 20 bp. For the two inserted binding sites we used the consensus sequences given by the position weight matrices V\$IRF1\_01 and V\$USF\_01, respectively.

The second data set is a breast cancer (BRC) gene set determined by Sorlie et al. (2003) and taken from Joshi et al. (2012). The genes have been identified based on their

differential mRNA expression behavior in cancer cells and are grouped according to their expression pattern into the five molecular breast cancer-associated subtypes: Luminal A, Luminal B, Normal-like, ErbB2 over-expressing and Basal-like using hierarchical clustering (Sorlie et al., 2003). Our analysis is based on the promoter sequences of the associated genes. The number of genes as well as their corresponding promoter sequences (−500 bp to +100 bp relative to the transcription start site defined by Joshi et al. (2012) in each subtype are given in **Table 4**. It can be seen that the BRC-subtype data sets differ in the number of genes and consequently in the number of promoter sequences. For example, *Luminal A* gene set appears to be the largest set by consisting of 86 promoter sequences and in turn, the set *ErbB2 over-expressing* is the smallest sequence set by owning 15 promoter sequences (see **Table 4**). Such differences are important and make it possible to demonstrate the functionality of our extension approach for different sequence-set sizes.

The Methods section of this study comprises two main parts. First, we review our previous work PC-TraFF (Meckbach et al., 2015) so that the readers have sufficient background information to understand the proposed extension in the PC-TraFF workflow. After that, we present our proposed extension approach for the separation of sequence-set specific TF cooperations from common (generally important) ones.

#### Previous Work: Introduction to PC-TraFF

PC-TraFF is an information theory based method that uses the pointwise mutual information (PMI) for the identification of potentially cooperating transcription factors according to their binding site pattern in a set of sequences. The algorithm of PC-TraFF comprises six phases and provides for each TFBS-pair  $t_a$  and  $t_b$  a  $\text{PMI}_{pc}(t_a, t_b)$ -value based on their distances and frequencies in the sequences, under study.

The overall workflow of PC-TraFF can be briefly given as:

##### 3.1.1. Phase 1: Construction and Filtering of the TFBS-Sequence Matrix

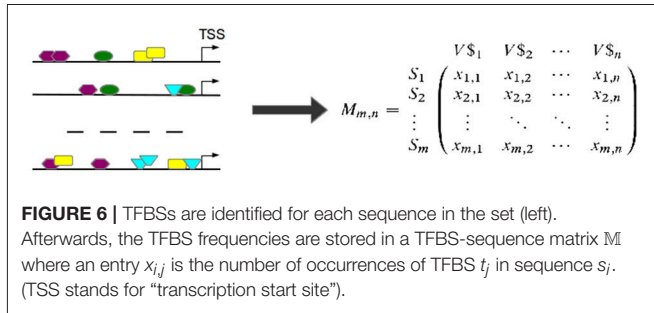
In the first step we predict all transcription factor binding sites (TFBSs) in a set of sequences by applying Match<sup>TM</sup> program (Kel et al., 2003) using the profile parameters and the position weight matrix (PWM) library specified in Deyneko et al. (2013). The PWMs are taken from TRANSFAC database (Wingender, 2008).





**TABLE 4 |** The number of genes and promoter sequences for the BRC-associated subtypes.

BRC subtypes	Number of genes	Number of promoter sequences
Luminal A	78	86
Luminal B	55	57
Normal-like	23	27
Basal-like	28	31
ErbB2 over-expressing	13	15



### 3.1.2. Phase 2: Identification of Important TFBSs in Each Sequence

In order to identify important TFBSs for each sequence, we calculate the pointwise mutual information  $\mathbb{PMI}(s_i; t_j)$  for each sequence  $s_i$  and TFBS  $t_j$  pair based on the frequencies of observed TFBSs in each sequence.

$$\mathbb{PMI}(s_i; t_j) = \log_2 \frac{p(s_i, t_j)}{p(s_i)p(t_j)},$$

where  $p(s_i, t_j)$  is the probability of a TFBS  $t_j$  to occur in sequence  $s_i$ . It is calculated as

$$p(s_i, t_j) = \frac{f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \tag{1}$$

where  $f_{ij}$  is the frequency of TFBS  $t_j$  in sequence  $s_i$ .  $p(s_i)$  and  $p(t_j)$  are the marginal probabilities and are calculated as

$$p(s_i) = \frac{\sum_{j=1}^n f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \tag{2}$$

A TFBS  $t_j$  is regarded to be important for sequence  $s_i$  if the corresponding  $\mathbb{PMI}(s_i, t_j) > 0$ . In the following analysis steps, for each sequence only the important TFBSs are considered.

### 3.1.3. Phase 3: Filter to Avoid Overlaps

Overlapping TFBSs of the same type are filtered in a way that the TFBS survives which is closer to TSS in order to avoid the overestimation of these repetitive binding sites (see **Figure 7A**) and thereby to consider only these TFBSs that appear to be more functional (Whitfield et al., 2012).

### 3.1.4. Phase 4: Construction of TFBS Pairs

TFBS pairs are identified according to the distance of their centers (see **Figure 7B**). Two TFBSs can form a pair if their distance satisfies the pre-defined minimal and maximal thresholds.

### 3.1.5. Phase 5: Weighted Cumulative Pointwise Mutual Information

The weighted cumulative pointwise mutual information  $\mathbb{PMI}_{pc}(t_a; t_b)$  of two putative TFBSs  $t_a$  and  $t_b$  is calculated as follows:

$$\mathbb{PMI}_{pc}(t_a; t_b) = \sum_{s \in S} w_s \cdot p(t_a, t_b) \cdot \log_2 \frac{p(t_a, t_b)}{p(t_a) \cdot p(t_b)}, \tag{3}$$

where  $p(t_a, t_b)$ ,  $p(t_a)$  and  $p(t_b)$  are the joint and marginal probabilities of TFBSs  $t_a$  and  $t_b$ , respectively. Further,  $w_s$  refers to the weight of a sequence  $s$  and is calculated based on the number of TFBS pairs  $N_s$  in  $s$  divided by the total number of TFBS pairs in the entire set of sequences  $S$ .

$$w_s = \frac{N_s}{\sum_{s_i \in S} N_{s_i}} \tag{4}$$

### 3.1.6. Phase 6: Background Noise Reduction of TFBSs Using Average Product Correction

To this end, using the average product correction (APC) theorem proposed by Dunn et al. (2008), the  $\mathbb{PMI}_{pc}(t_a; t_b)$  scores have been adjusted:

$$\mathbb{PMI}_{pc}^{APC}(t_a; t_b) = \mathbb{PMI}_{pc}(t_a; t_b) - \frac{\mathbb{PMI}_{pc}(t_a; \bar{t}_x) \cdot \mathbb{PMI}_{pc}(t_b; \bar{t}_x)}{\bar{\mathbb{PMI}}_{pc}} \tag{5}$$

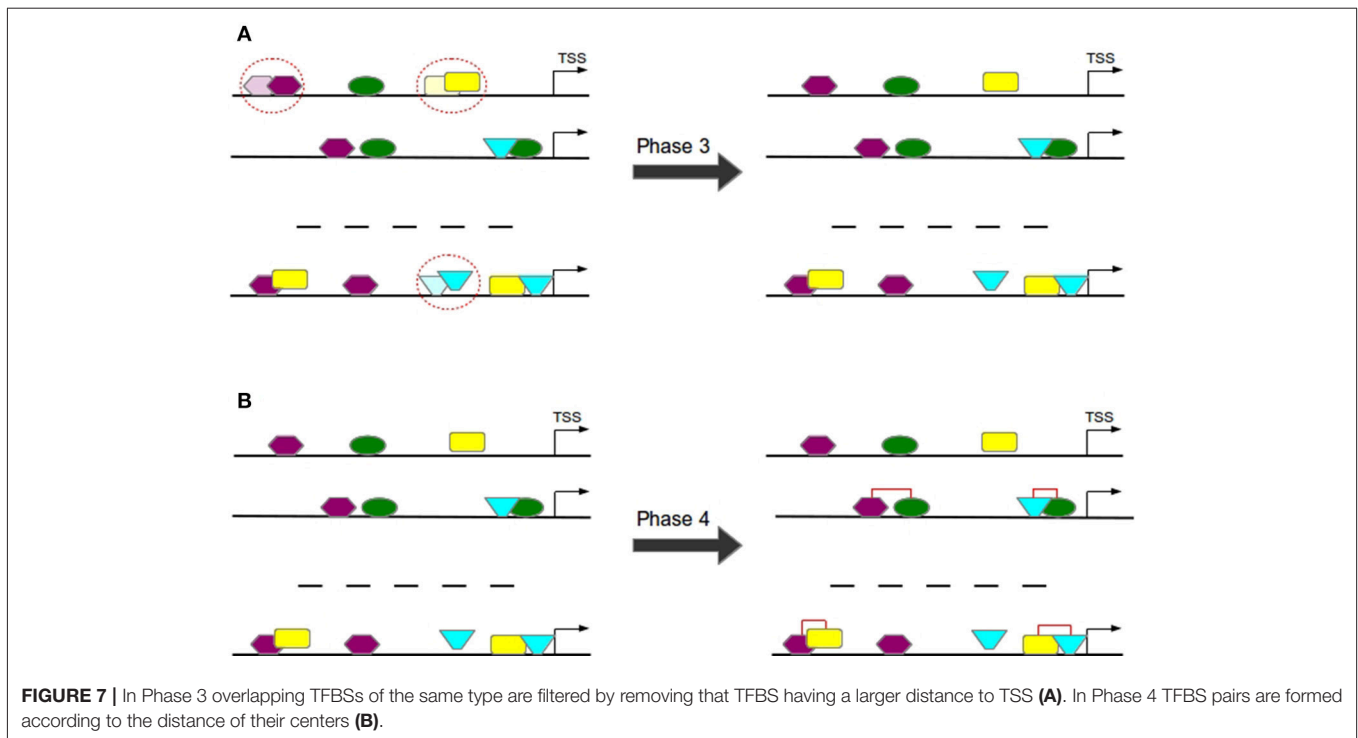
where  $\mathbb{PMI}_{pc}(t_a; \bar{t}_x)$  is the mean  $\mathbb{PMI}_{pc}$  of  $t_a$  to all other TFBSs in the sequences, and  $\bar{\mathbb{PMI}}_{pc}$  is the mean  $\mathbb{PMI}_{pc}$  value over all TFBS pairs.

The resulting  $\mathbb{PMI}_{pc}^{APC}$  values are transformed into z-scores and only those pairs are considered to be significant that have a z-score  $\geq 3$ .

### Separation of Sequence Set Specific TF Cooperations From the Common Ones

According to their TFBS motifs, some TF cooperations are noticeable sensitive to the context of nucleotides - regarding the order and positions of nucleotides in sequences - in comparison to common TF cooperations, which are often found as significant for different sequence sets.

In order to separate such sequence-set specific significant TFBS pairs from the common (general important) significant pairs, we propose the following approach: The uShuffle algorithm (Jiang et al., 2008) is used to shuffle the nucleotides within each sequence by setting k-mers' size = 3. Thereby, not only the single nucleotide counts of each sequence are maintained but also the triplet counts and thus, the core of TFBSs. By repeating this



shuffling process several times, a sufficient number of randomly generated sequence sets (e.g., 1000) is created.

Second, employing the Match<sup>TM</sup> algorithm for each set of shuffled sequences, the putative binding sites of TFs in these sequences are predicted. Third, applying PC-TraFF algorithm, new  $\mathbb{P}MII_{pc}$ -values for every TFBS pair in each randomly generated sequence set are calculated. Fourth, based on these  $\mathbb{P}MII_{pc}$ -values of each pair  $t_a$  and  $t_b$ , we define the average  $\mathbb{P}MII$ -value,  $AVG(\mathbb{P}MII(t_a; t_b))$  as

$$AVG(\mathbb{P}MII(t_a; t_b)) = \frac{1}{l} \sum_{i=1}^l \mathbb{P}MII_{pc}^{APC}(t_a; t_b)_i, \quad (6)$$

where  $l$  is the number of randomly generated sequence sets.

After that, the  $AVG(\mathbb{P}MII(t_a; t_b))$ -value of binding sites  $t_a$  and  $t_b$  is subtracted from their initial significant  $\mathbb{P}MII_{pc}^{APC}(t_a; t_b)$ -value as

$$\mathbb{P}MII^{specific}(t_a; t_b) = \mathbb{P}MII_{pc}^{APC}(t_a; t_b) - \left[ (1 + \alpha) \times AVG(\mathbb{P}MII(t_a; t_b)) \right], \quad (7)$$

where  $\alpha \in [-1, +1]$  is a preassigned real number for monitoring the influence of this process on the significant TFBS pairs. It can easily be seen that  $\alpha = -1$  results in the original PC-TraFF analysis. By setting  $\alpha = 0$  the average  $AVG(\mathbb{P}MII(t_a; t_b))$  is subtracted from the original  $\mathbb{P}MII_{pc}^{APC}(t_a; t_b)$  value whereas an  $\alpha \geq 0$  leads to a stronger effect of the subtraction and thus, a more strict selection process. However, for the proper application of this process the determination of an upper bound

for  $\alpha$  is crucial in order to avoid the overestimation of the efficacy of  $AVG(\mathbb{P}MII(t_a; t_b))$ -values (background level) on the separation of sequence-set specific pairs from common ones. By systematically analyzing different values, we established that  $+1$  is the most convenient upper bound for  $\alpha$ .

A positive  $\mathbb{P}MII^{specific}(t_a; t_b)$ -value of binding sites  $t_a$  and  $t_b$  identified in the promoter sequences of a certain sequence set suggests that the binding of the related TF pair is strongly sequence context dependent. In contrast, a  $\mathbb{P}MII^{specific}(t_a; t_b)$ -value  $\leq 0$  indicates that the cooperations of corresponding TFs could have a general importance for the controlling of genetic programs.

## 4. CONCLUSIONS

Depending on their biological functions as well as cellular context, TFs specify the selection of cooperation partners in many ways for different cell types. However, the existing algorithms often focus on the identification of all predictable TF cooperations without distinguishing between sequence-set specific and common, i.e., ubiquitously occurring TF cooperations. To address this limitation, we propose in this study an approach that extends our previous method PC-TraFF in order to assign its predictions into two main categories: sequence-set specific and common (generally important) ones. For this aim, we estimated the background co-occurrence of any TF pair by preserving the nucleotide composition and the core of TFBS motifs in the sequences of interest. To maintain the core of TFBS motifs, we set the  $k$ -mers'size = 3 in the randomly shuffled new sets of sequences. It can be seen that,

while an increase in  $k$ -mers' size could lead to increment of background co-occurrence of TFBSs, a decrease in  $k$ -mers' size could in turn result in the reduction of background level of TF pairs. In order to assess the effectiveness of our extension approach, we analyzed promoter sequences of five different breast cancer-associated subtypes. The results show that the cooperating pairs identified by original PC-TraFF algorithm were considerably overlapping between the subtypes. Applying our extension approach, we could successfully separate sequence-set specific pairs from common ones and thereby reducing the number of overlapping pairs. Further, when we applied our extension approach of the original PC-TraFF algorithm to a simulation data set with varying  $\alpha$ -values and, thus, different background levels, we could demonstrate that the cooperating TF pair was consistently identified as a sequence-set specific pair. The scaling parameter  $\alpha$  is useful to extend or reduce the level of the subtracted background. Thereby, the influence of  $\alpha$  itself is not linear but highly depending on the sequence set and thus on the respective background. Starting with an  $\alpha$ -value of 0.2 we recommend to slightly increase  $\alpha$  in order to assess the effect of  $\alpha$  on the given data set and in doing so, to get the desired ratio between sensitivity and specificity. In summary, the proposed extension approach can successfully be applied for the distinction of sequence-set specific TF cooperations from common ones which are identified as generally important for different data sets.

## REFERENCES

- Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956. doi: 10.1016/j.cell.2005.08.020
- Chatr-aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., et al. (2014). The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 43, D470–D478. doi: 10.1093/nar/gku1204
- Chuang, C.-L., Hung, K., Chen, C.-M., and Shieh, G. S. (2009). Uncovering transcriptional interactions via an adaptive fuzzy logic approach. *BMC Bioinformatics* 10:400. doi: 10.1186/1471-2105-10-400
- Conway, J. R., Lex, A., and Gehlenborg, N. (2017). Upsetr: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364
- Deyneko, I., Kel, A., Kel-Margoulis, O., Deineko, E., Wingender, E., and Weiss, S. (2013). MatrixCatch - a novel tool for the recognition of composite regulatory elements in promoters. *BMC Bioinformatics* 14:241. doi: 10.1186/1471-2105-14-241
- Dongen, S. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, Netherlands.
- Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24, 333–340. doi: 10.1093/bioinformatics/btm604
- Girgis, H., and Ovcharenko, I. (2012). Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. *BMC Bioinformatics* 13:25. doi: 10.1186/1471-2105-13-25
- Ha, N., Polychronidou, M., and Lohmann, I. (2012). COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. *PLoS ONE* 7:e52055. doi: 10.1371/journal.pone.0052055
- Hu, Z., and Gallo, S. M. (2010). Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics* 11:49. doi: 10.1186/1471-2164-11-49
- Hu, Z., Hu, B., and Collins, J. (2007). Prediction of synergistic transcription factors by function conservation. *Genome Biol.* 8:R257. doi: 10.1186/gb-2007-8-12-r257
- Jankowski, A., Prabhakar, S., and Tiurnyn, J. (2014). TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics* 15:208. doi: 10.1186/1471-2164-15-208
- Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics* 9:192. doi: 10.1186/1471-2105-9-192
- Joshi, H., Nord, S. H., Frigessi, A., Børresen-Dale, A.-L., and Kristensen, V. N. (2012). Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes. *BMC Genomics* 13:199. doi: 10.1186/1471-2164-13-199
- Kel, A., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E. (2003). MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31, 3576–3579. doi: 10.1093/nar/gkg585
- Kel-Margoulis, O., Kel, A., Reuter, I., Deineko, I., and Wingender, E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* 30, 332–334. doi: 10.1093/nar/30.1.332
- Lai, F.-J., Jhu, M.-H., Chiu, C.-C., Huang, Y.-M., and Wu, W.-S. (2014). Identifying cooperative transcription factors in yeast using multiple data sources. *BMC Syst. Biol.* 8:S2. doi: 10.1186/1752-0509-8-S2-S2
- Meckbach, C., Tacke, R., Hua, X., Waack, S., Wingender, E., and Gültas, M. (2015). PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinformatics* 16:400. doi: 10.1186/s12859-015-0827-2
- Mysickova, A., and Vingron, M. (2012). Detection of interacting transcription factors in human tissues using predicted DNA binding affinity. *BMC Genomics* 13(Suppl 1):S2. doi: 10.1186/1471-2164-13-S1-S2
- Nandi, S., Blais, A., and Ioshikhes, I. (2013). Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Res.* 41, 8822–8841. doi: 10.1093/nar/gkt578
- Navarro, C., Lopez, F. J., Cano, C., Garcia-Alcalde, F., and Blanco, A. (2014). CisMiner: Genome-wide *in-Silico* cis-regulatory module prediction by fuzzy itemset mining. *PLoS ONE* 9:e108065. doi: 10.1371/journal.pone.0108065

## AVAILABILITY OF DATA AND ALGORITHM

The extension of PC-TraFF is freely accessible at <http://pctraffpro.bioinf.med.uni-goettingen.de/>. All data sets and results of this paper are available from the corresponding author on request.

## AUTHOR CONTRIBUTIONS

CM and MG developed the model and conducted computational analyses. EW interpreted the results and adjusted the model together with CM and MG. CM and MG conceived of and managed the project and wrote the final version of the manuscript. All authors read and approved the final manuscript.

## FUNDING

CM was funded by ExiTox2 (Förder Kennzeichen: 031L0120B) of the BMBF (German Ministry of Education and Research).

## ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

- Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274–1286. doi: 10.1016/j.cell.2012.04.040
- Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8418–8423. doi: 10.1073/pnas.0932692100
- Spadafore, M., Najarian, K., and Boyle, A. P. (2017). A proximity-based graph clustering method for the identification and application of transcription factor clusters. *BMC Bioinformatics* 18:530. doi: 10.1186/s12859-017-1935-y
- Sun, H., Guns, T., Fierro, A. C., Thorrez, L., Nijssen, S., and Marchal, K. (2012). Unveiling combinatorial regulation through the combination of ChIP information and *in silico* cis-regulatory module detection. *Nucleic Acids Res.* 40:e90. doi: 10.1093/nar/gks237
- Teif, V. B., and Rippe, K. (2010). Statistical-mechanical lattice models for protein-DNA binding in chromatin. *J. Phys. Condens Matter* 22:414105. doi: 10.1088/0953-8984/22/41/414105
- Van Loo, P., and Marynen, P. (2009). Computational methods for the detection of cis-regulatory modules. *Brief. Bioinform.* 10, 509–524. doi: 10.1093/bib/bbp025
- Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., et al. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 13:R50. doi: 10.1186/gb-2012-13-9-r50
- Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* 9, 326–332. doi: 10.1093/bib/bbn016
- Wu, W.-S., and Lai, F.-J. (2016). Detecting cooperativity between transcription factors based on functional coherence and similarity of their target gene sets. *PLoS ONE* 11:e0162931. doi: 10.1371/journal.pone.0162931
- Zeidler, S., Meckbach, C., Tacke, R., Raad, F., Roa, A., Uchida, S., et al. (2016). Computational detection of stage-specific transcription factor clusters during heart development. *Front. Genet.* 7:33. doi: 10.3389/fgene.2016.00033

**Conflict of Interest Statement:** EW is head of geneXplain GmbH, the company that maintains and distributes the TRANSFAC database.

The other authors declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Meckbach, Wingender and Gültas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

### **A.3. Computational detection of stage-specific TF clusters during heart development**





# Computational Detection of Stage-Specific Transcription Factor Clusters during Heart Development

Sebastian Zeidler<sup>1,2,3\*</sup>, Cornelia Meckbach<sup>1</sup>, Rebecca Tacke<sup>1</sup>, Farah S. Raad<sup>2,3</sup>, Angelica Roa<sup>2,3</sup>, Shizuka Uchida<sup>4,5</sup>, Wolfram-Hubertus Zimmermann<sup>2,3</sup>, Edgar Wingender<sup>1,3</sup> and Mehmet Gültas<sup>1</sup>

<sup>1</sup> University Medical Center Göttingen, Institute of Bioinformatics, Georg-August-University Göttingen, Göttingen, Germany, <sup>2</sup> Heart Research Center Göttingen, University Medical Center Göttingen, Institute of Pharmacology and Toxicology, Georg-August-University Göttingen, Göttingen, Germany, <sup>3</sup> DZHK (German Centre for Cardiovascular Research), Göttingen, Germany, <sup>4</sup> Institute of Cardiovascular Regeneration, Goethe University Frankfurt, Frankfurt, Germany, <sup>5</sup> DZHK (German Centre for Cardiovascular Research), Frankfurt, Germany

## OPEN ACCESS

### Edited by:

Yasset Perez-Riverol,  
European Bioinformatics Institute, UK

### Reviewed by:

Mikhail P. Ponomarenko,  
Institute of Cytology and Genetics of  
Siberian Branch of Russian Academy  
of Sciences, Russia  
Ka-Chun Wong,  
City University of Hong Kong, China

### \*Correspondence:

Sebastian Zeidler  
sebastian.zeidler@  
bioinf.med.uni-goettingen.de

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 05 November 2015

**Accepted:** 23 February 2016

**Published:** 23 March 2016

### Citation:

Zeidler S, Meckbach C, Tacke R, Raad FS, Roa A, Uchida S, Zimmermann WH, Wingender E and Gültas M (2016) Computational Detection of Stage-Specific Transcription Factor Clusters during Heart Development. *Front. Genet.* 7:33. doi: 10.3389/fgene.2016.00033

Transcription factors (TFs) regulate gene expression in living organisms. In higher organisms, TFs often interact in non-random combinations with each other to control gene transcription. Understanding the interactions is key to decipher mechanisms underlying tissue development. The aim of this study was to analyze co-occurring transcription factor binding sites (TFBSs) in a time series dataset from a new cell-culture model of human heart muscle development in order to identify common as well as specific co-occurring TFBS pairs in the promoter regions of regulated genes which can be essential to enhance cardiac tissue developmental processes. To this end, we separated available RNAseq dataset into five temporally defined groups: (i) mesoderm induction stage; (ii) early cardiac specification stage; (iii) late cardiac specification stage; (iv) early cardiac maturation stage; (v) late cardiac maturation stage, where each of these stages is characterized by unique differentially expressed genes (DEGs). To identify TFBS pairs for each stage, we applied the MatrixCatch algorithm, which is a successful method to deduce experimentally described TFBS pairs in the promoters of the DEGs. Although DEGs in each stage are distinct, our results show that the TFBS pair networks predicted by MatrixCatch for all stages are quite similar. Thus, we extend the results of MatrixCatch utilizing a Markov clustering algorithm (MCL) to perform network analysis. Using our extended approach, we are able to separate the TFBS pair networks in several clusters to highlight stage-specific co-occurrences between TFBSs. Our approach has revealed clusters that are either common (NFAT or HMG1Y clusters) or specific (SMAD or AP-1 clusters) for the individual stages. Several of these clusters are likely to play an important role during the cardiomyogenesis. Further, we have shown that the related TFs of TFBSs in the clusters indicate potential synergistic or antagonistic interactions to switch between different stages. Additionally, our results suggest that cardiomyogenesis follows the hourglass model which was already proven for *Arabidopsis* and some vertebrates. This investigation helps us to get a better understanding of how each stage of cardiomyogenesis is affected by different combination of TFs. Such knowledge may help to understand basic principles of stem cell differentiation into cardiomyocytes.

**Keywords:** cardiomyogenesis, engineered heart muscle, MatrixCatch, Markov clustering, transcription factor collaboration

## 1. INTRODUCTION

Transcription factors (TFs) regulate the expression of genes and genetic programs to maintain survival and adaptation to the environment in adult organisms as well as in embryonic and organogenesis. Most of them bind to recognized specific sequences in the DNA regulatory regions of genes and modify transcription, such as the assembly of the gene expression machinery. In mammalian tissues TFs often work in combinatorial interactions for precise regulation of specific programs (Boyer et al., 2005; Odom et al., 2006; Hu and Gallo, 2010; Neph et al., 2012). Such interactions can be positive, resulting in an enhanced expression of a gene or negative, resulting in reduced expression of a target gene. Thus, the identification of co-occurring transcription factor binding sites (TFBSs) in the promoter regions of regulated genes indicate potential combinatorial interactions between TFs that are important for understanding the molecular mechanisms, e.g., of tissue development during embryogenesis.

The human heart is the first organ formed during embryogenesis (Kirby, 2002; Brand, 2003; Buckingham et al., 2005; Brewer and Pizzey, 2006; Schleich et al., 2013), and it consists of different cell types, which develop simultaneously and are regulated by TFs as well as their combinatorial interactions. Until now, several groups analyzed TFs and their influence on cardiac development (Ryan and Chin, 2003; Pikkarainen et al., 2004; Peterkin et al., 2005; Brewer and Pizzey, 2006; Martin et al., 2010; Shi and Jin, 2010; Turbendian et al., 2013; Chaudhry et al., 2014; Takeuchi, 2014; Wang and Jauch, 2014). These studies mainly focus on individual TFs or their related families e.g., GATA family, TBX family, or NKX2 family (Ryan and Chin, 2003; Pikkarainen et al., 2004; Miura and Yelon, 2013; Turbendian et al., 2013). However, a detailed analysis of interactions between TFs and their role in cardiac development is limited to interactions between known cardiac TFs like NKX2-5 or MEF2 which are essential for the generation of cardiac tissues from stem cells (Martin et al., 2010; Sylva et al., 2014; Takeuchi, 2014). A complete survey of potential TF interactions by co-occurring TFBSs in the promoter regions of genes which regulate cardiac development is still missing, but needed to understand embryonic cardiac development, in particular of cardiomyocytes (CMs).

CMs comprise the most important functional cells in the human heart (Ye et al., 2013; Sylva et al., 2014). CMs show a limited potential to regenerate after myocardial infarction or other cardiovascular diseases (CVDs), which is at maximum 50% CM renewal per lifetime and less than 1% per year (Bergmann et al., 2009; Sylva et al., 2014; Takeuchi, 2014). Replacing CMs in elderly by for example enhanced cardiomyocyte proliferation may improve the quality of their life, but requires an understanding of how CMs develop and of how they can be replaced (Akhurst, 2012; Ye et al., 2013; Euler, 2015).

One approach is to apply tissue engineered myocardium to restore muscle mass and thus reintroduce contractility (Zimmermann et al., 2006). Such tissues can be generated from embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs), or parthenogenetic stem cells (Soong et al., 2012; Didié

et al., 2013; Ye et al., 2013; Tiburcy and Zimmermann, 2014). Controlling cardiomyogenesis *in vitro* requires insight into biological processes governing embryonic heart development. To understand cardiac development from a systems biology perspective, identification of the mechanisms controlling the expression of fate determining TFs and their regulation of transcription are of fundamental importance. Co-occurring TFBSs in the regulatory regions of genes which are specific for a particular developmental stage reveal potential TF interactions that are likely to regulate these stages. There are in fact plenty of TF-TF interactions known as implicated in organogenesis, but the specific time points when particular interactions occur, are difficult to obtain and mostly not annotated in public databases. Only intense literature surveys provide such information.

Recent studies identifying the co-occurrence of TF pairs focus either on combinatorial approaches where e.g., specific DNA-sequences bound by different TFs simultaneously were selected from a library of random sequences (Jolma et al., 2015) or approaches that focus on data integration e.g., ChIP-seq, SELEX together with Hi-C to reveal long-range chromatin interactions (Jolma et al., 2013; Wong et al., 2016). Although the selection of interacting TF pairs from a library of random sequences underpins potential interactions of TFs, it does not give any hints on the actual interactions in particular cell types or tissues. Data integration and especially Hi-C technology is very promising for the future, but currently there is a lack in publicly available data sets that cover the time dependent organogenesis of the human heart.

In this study we analyze a time series dataset obtained from RNAseq at different time points of *in vitro* cardiomyogenesis (Hudson et al.; in revision) to identify co-occurring TFBSs which indicate potential interacting TFs that are crucial for understanding the gene regulatory mechanisms during the heart development. The dataset consists of six different time points (day: 0, 3, 8, 13, 29, and 60) where the gene expression in the tissue culture was measured by RNAseq. The data comprises early heart development in general and can be differentiated in the following major developmental stages: (i) mesoderm induction stage (day 0–day 3); (ii) cardiac specification stage (day 3–day 13; early 3–8, late 8–13); (iii) cardiac maturation stage (day 13–day 60; early 13–29, late 29–60). For each stage we determined the set of unique differentially expressed genes (DEGs) utilizing *limma* on the FPKM-values in the dataset (Smyth, 2004). To identify specific TF interactions in individual stages, we analyzed the promoter sequences of corresponding DEGs employing the MatrixCatch approach (Deyneko et al., 2013). As a result, we observed a set of co-occurring TFBSs for each stage whose corresponding TFs are likely to represent potential core regulators of a particular developmental stage. Although the analyzed DEGs are unique in each stage, the identified TFBS pairs are highly overlapping between stages. To overcome this problem in MatrixCatch results, we further applied Markov clustering algorithm (MCL; Dongen, 2000) for the detection of clusters which contain stage specific co-occurrences between TFBSs. In recent years, MCL has gained great attention in the bioinformatics community for the detection of high-quality clusters in biological networks due to its highly effective



and successful algorithm. Especially, for the clustering of protein-protein interaction networks, several studies have shown that MCL is superior to conventional clustering approaches in terms of detection of high-quality and more accurate functional clusters (Brohée and van Helden, 2006; Vlasblom and Wodak, 2009; Shih and Parthasarathy, 2012). These articles encouraged us to utilize MCL for the elimination of negligible pairs at each stage and thus for the determination of remaining TFBS pairs, which may play crucial roles during cardiomyogenesis. To this end, we focused on clusters whose central binding site is present at almost all stages, but its partners differ stage-specifically. These clusters may regulate DEGs in each stage and are likely to be fundamentally implicated in cardiac muscle development.

## 2. MATERIALS AND METHODS

In this section we describe the differentially expressed genes analyzed and the methods applied and partly developed. Our analysis follows the structure of **Figure 1**.

### 2.1. Selection of Differentially Expressed Genes

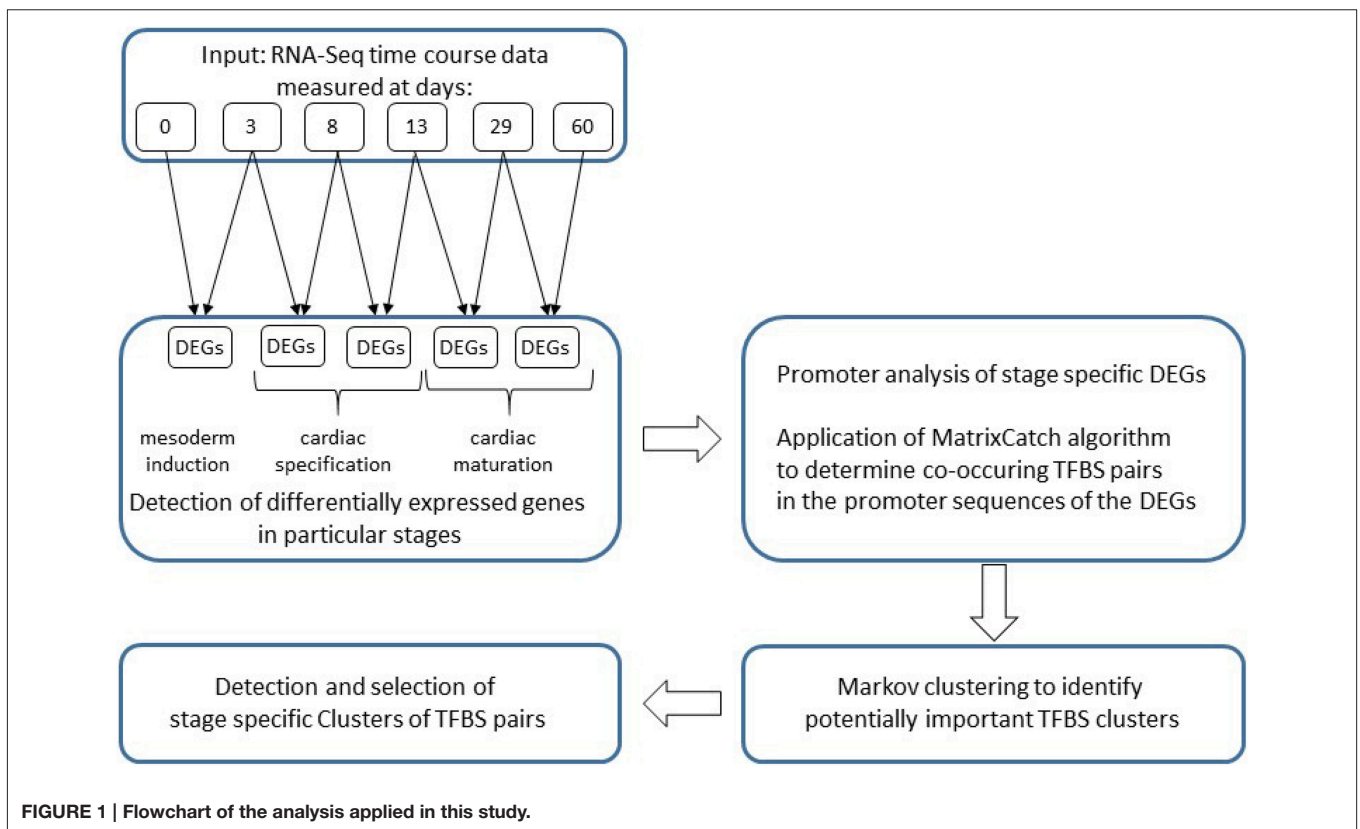
The data, available as a FPKM normalized RNAseq time series, was mapped to corresponding gene symbols (hgnc-symbols) and further analyzed using *limma* package from the Bioconductor project for R with standard procedures (Smyth, 2004; R Core Team, 2015). The time series data describe human

cardiomyogenesis in vitro at time points day 0, 3, 8, 13, 29, and 60, whereas day 0 resembles blastocyst stage development and day 60 early fetal stages (Hudson et al.; in revision). We calculated DEGs between two time points which define a particular developmental stage where: (i) day 0–3 defines the mesoderm induction stage; (ii) day 3–8 early cardiac specification; (iii) day 8–13 late cardiac specification; (iv) day 13–29 early cardiac maturation and; (v) day 29–60 the late cardiac maturation stage (this stage describes the transition from an embryonic to a fetal cardiac maturation stage). We filtered the set of all DEGs for protein coding genes (excluding TFs) and their uniqueness in a stage by comparison to all other stages with  $p$ -value  $\leq 0.05$  and FDR  $\leq 0.01$  (see **Supplementary File 1**). A heatmap of stage-specific DEGs is given in **Supplementary File 2**.

### 2.2. Promoter Sequences

Using UCSC genome browser (Karolchik et al., 2004), we extracted for each protein coding gene (RefSeq gene) based on its annotated transcription start site (TSS) the -1 kb putative regulatory promoter region.

It is important to note that, according to TSS annotations, a RefSeq gene can have multiple overlapping promoter regions which results in overestimation of the importance of some transcription factor binding sites (TFBSs). Thus, following the line of PC-TraFF to remove the redundancy between sequences, we filtered them regarding their TSSs (Meckbach et al., 2015). Consequently, we used in our analysis only those sequences which have no overlap.



**FIGURE 1 |** Flowchart of the analysis applied in this study.

In this study, the assembly of the hg19 release of the human genome was used and only UCSC track refGene annotations were considered which correspond to the chromosomes chr1-chr22, chrX, and chrY.

### 2.3. MatrixCatch Analysis

MatrixCatch is a novel method introduced by Deyneko et al. (2013) to recognize experimentally verified TF pairs based on the co-localization of their TFBSs, known as composite regulatory modulues (CRMs), in single promoters. To detect CRMs in the individual sequences under study, MatrixCatch scans each sequence and its reverse complement using a special library of position weight matrices (PWMs). This library has been specified by considering the TF binding scores, relative orientations and distances between TFs that are experimentally known to interact, as documented in the TRANSCompel database (Kel-Margoulis et al., 2002). Consequently, the usage of MatrixCatch yields an important practical advantage since this method provides a high number of known CRMs in sequences with their biological interpretation (for details, see Deyneko et al., 2013).

In our study, we applied MatrixCatch to the promoter sequences of the filtered DEGs of the different heart developmental stages. As we have recently suggested in PC-TraFF (Meckbach et al., 2015), we prefer in this study the usage of TFBS pairs instead of CRMs, since those pairs were detected in a set of sequences. This indicates the importance of potential collaborations between corresponding TFs in the gene set of interest.

### 2.4. Clustering of Co-Occurring TFBSs

Since MatrixCatch provides all detected TFBS pairs of experimentally verified TF interactions in promoters, the detected pairs are highly overlapping between developmental stages. To differentiate stage specific roles of TFBS pairs, we first determined the frequency of each pair in MatrixCatch results. After that, we applied the Markov clustering algorithm (MCL; Dongen, 2000) which is able to eliminate negligible TFBS pairs based on their frequencies at each stage. To this end, we constructed an interaction network based on the TFBS pairs for each heart developmental stage, where nodes are TFBSs and edges display the co-occurrences between them.

Let  $\mathcal{N} = (\mathcal{V}, \mathcal{E})$  be an undirected interaction network of TFBS pairs where any two elements  $(v_i, v_j \in \mathcal{V})$  of  $\mathcal{N}$  are connected by an edge  $e_{(v_i, v_j)}$  belonging to  $\mathcal{E}$ , if and only if the corresponding TFBS pair was identified by MatrixCatch. Further,  $w(v_i, v_j)$  denotes the weight of an edge  $e_{(v_i, v_j)}$ , which represents the observed frequency of the TFBS pair  $(v_i, v_j)$  found by MatrixCatch in the promoter sequences of genes under study.

Based on the weights of edges, an adjacency matrix  $\mathcal{A}_{n \times n}$  of each network was constructed as

$$A_{i,j} = \begin{cases} w(v_i, v_j) & \text{if } e_{(v_i, v_j)} \in \mathcal{E} \\ 0 & \text{else.} \end{cases}$$

$\mathcal{A}_{n \times n}$  was then converted into a row stochastic "Markov" matrix  $\mathcal{M}_{n \times n}$ , where  $m_{i \times j}$  represents the transition probability between nodes  $v_i$  and  $v_j$  in the network under study. The most common

way to construct a row stochastic transition matrix  $\mathcal{M}$  is the normalization of rows in  $\mathcal{A}$  to sum to 1. This process can be simply given as:  $\mathcal{M} = \Delta^{-1} \cdot \mathcal{A}$ , where  $\Delta$  is a  $n \times n$  diagonal degree matrix and defined as:

$$\Delta = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^n a_{1j} & 0 & \dots & 0 \\ 0 & \sum_{j=1}^n a_{2j} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sum_{j=1}^n a_{nj} \end{pmatrix}$$

Based on matrix  $\mathcal{M}$ , we employed MCL (Dongen, 2000) to detect densely connected TFBSs in each network. Briefly, the basic intuition of MCL was based on a simulation of stochastic flows on the underlying interaction network to separate high-flow regions from low-flow regions. To this end, Expand and Inflate operations were applied on  $\mathcal{M}$  until  $\mathcal{M}$  reaches its steady state. While the Expand operation corresponds to matrix multiplication ( $\mathcal{M} = \mathcal{M} \times \mathcal{M}$ ), the Inflate operation is used to increase the contrast between higher and lower probability transitions by taking each entry  $m_{i \times j}$  in  $\mathcal{M}$  to the power of inflation parameter  $r > 1$ . Finally,  $\mathcal{M}$  was re-normalized into a row stochastic matrix. The pseudo-code for MCL is given in Algorithm 1.

---

#### Algorithm 1 : Markov Clustering Algorithm

---

**Input:**  $\mathcal{M}$  and  $r > 1$

**Output:**  $\mathcal{C}$ : A list of clusters

**Method:**

- 1:  $t = 0$
  - 2:  $\mathcal{M}_t = \mathcal{M}$
  - 3: **repeat**
  - 4:      $t = t + 1$
  - 5:      $\mathcal{M}_t = \text{Expand}(\mathcal{M}_{t-1}) = \mathcal{M}_{t-1} \times \mathcal{M}_{t-1}$
  - 6:      $\mathcal{M}_t = \text{Inflate}(\mathcal{M}_t, r) = \left\{ \frac{(m_{ij})^r}{\sum_{k=1}^n (m_{ik})^r} \right\}_{i,j=1}^n$
  - 7: **until**  $\mathcal{M}_t$  converges
  - 8:  $\mathcal{C}$ : clusters( $\mathcal{M}_t$ )
- 

## 3. RESULTS

We analyzed a time course data set which covers heart muscle development in human embryonic stem cell derived tissue cultures at days 0, 3, 8, 13, 29, and 60 (Hudson et al., in revision). These time points cover the mesoderm induction stage (day 0–day 3), the cardiac specification stage (day 3–day 13), and the cardiac maturation stage (day 13–day 29). We further defined cardiac specification and cardiac maturation into two more stages, i.e.: (i) early cardiac specification and maturation stage from days 3–8 and days 13–29, respectively; (ii) late cardiac specification and maturation with transition from embryonic to fetal stages defined by culture days 8–13 and days 29–60, respectively. By comparison of neighboring time points, for each stage, we determined the set of DEGs and filtered them according to their uniqueness in a particular stage. Afterwards, we utilized

MatrixCatch to identify co-occurring pairs of TFBSs in the promoter regions of these DEGs. Consequently, we identified: (i) 63 TFBS pairs based on 429 DEGs for the mesoderm induction stage; (ii) 82 TFBS pairs based on 1233 DEGs for the early cardiac specification stage; (iii) 24 TFBS pairs based on 36 DEGs for the late cardiac specification stage; (iv) 52 TFBS pairs based on 205 DEGs for the early cardiac maturation stage; (v) 76 TFBS pairs based on 964 DEGs for the late cardiac maturation stage (see **Supplementary File 3**).

Due to underlying methodology of MatrixCatch, the detected TFBS pairs show a large overlap between different stages although they may play different roles in these stages. To reduce this drawback of MatrixCatch, we further applied Markov clustering algorithm that seeks to remove negligible TFBS pairs by emphasizing the roles of remaining pairs at each stage. Consequently, we obtained (i) 19 clusters for the mesoderm induction stage; (ii) 25 clusters for the early cardiac specification stage; (iii) 11 clusters for the late cardiac specification stage; (iv) 21 clusters for the early cardiac maturation stage, and (v) 24 clusters for the late cardiac maturation stage (see **Supplementary File 4**).

We focused only on clusters with V\$AP1\_01, V\$HMG1Y\_Q6, V\$SMAD\_Q6\_01, and V\$NFAT\_Q6 binding sites in their center (see **Figure 2**), because these clusters contain at least three interactions and the changes in their constitution provide crucial information about different cardiac developmental stages. We analyzed the TFBS pairs in these clusters according to their potential role in cardiac development. We omitted clusters, when the expression values of TF genes are below a certain threshold or their importance in heart development is currently unknown. For our analysis, we applied a FPKM threshold value of 10, which discriminates robustly between expressed TF genes and low or not expressed TF genes.

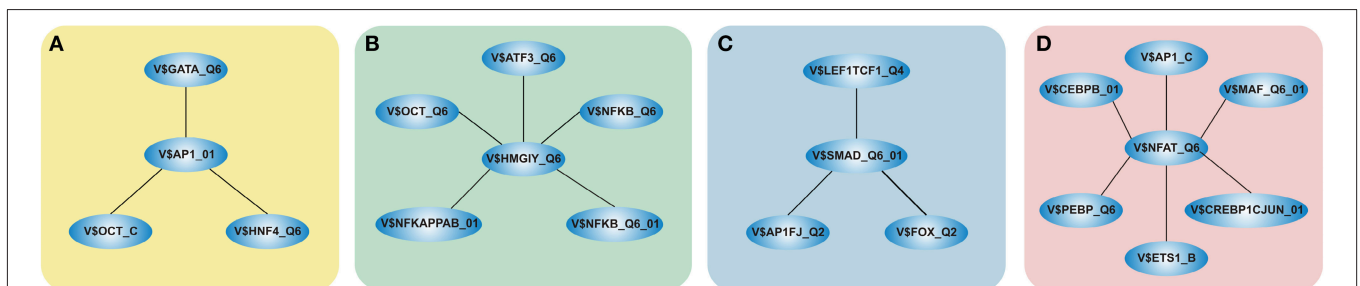
### 3.1. AP-1-Cluster

The AP-1-cluster is an assembly of different TFBSs with the V\$AP1\_01 binding site in its center (see **Figure 2A**). As described in **Table 1** and in **Figure 3**, V\$AP1\_01 binding site co-occurs with V\$OCT\_C binding site during mesoderm induction (< day 3) and early cardiac specification stage (day 3–day 8) and at late cardiac maturation stage (> day 29). Further, V\$AP1\_01 co-occurs with V\$GATA\_Q6 binding site at all stages except days

8–13. Interestingly, a co-occurring pair between V\$AP1\_01 and V\$HNF4\_Q6 binding site was detected only between day 3 and day 8. Additionally, **Figure 3** shows for these TFBSs the related TF genes which are expressed in at least one time point.

AP-1 is a family of leucine zipper transcription factors (bZIP) which forms homo- or heterodimers composed of proteins belonging to JUN or FOS protein families (Shaulian and Karin, 2002; Hess et al., 2004; Shaulian, 2010). AP-1 plays a role in the regulation of general functions like proliferation, differentiation, and apoptosis. We identified that V\$AP1\_01 co-occurs with V\$OCT\_C binding sites which are bound by AP-1 and POU-domain factors like POU5F1, respectively. POU5F1 is also known as OCT-4, which is an important pluripotency maintenance factor (Schöler et al., 1990; Nichols et al., 1998; Pesce and Schöler, 2001; Guo et al., 2002). Regarding the expression values, POU5F1 shows higher expression in early stages (< day 8) and is absent after day 13 (see **Figure 4B**). This is in contrast to AP-1, where AP-1 components (FOS as well as JUN) are not present or only present at reduced levels during early stages, but they show increased expression values after day 13 (see **Figure 4A**). This suggests that AP-1 may not be formed during early stages, where POU5F1 controls the associated genes, and that during the late cardiac maturation stage (> day 29) the analyzed genes are under control of AP-1.

Our analysis identified a co-occurrence of V\$AP1\_01 with V\$GATA\_Q6 binding sites. GATA factors form a protein family of six zinc finger transcription factors that share a highly conserved DNA-binding sequence (Orkin, 1992; Ohneda and Yamamoto, 2002; Pikkarainen et al., 2004; Brewer and Pizzezy, 2006). As suggested in Brewer and Pizzezy (2006), the family can be dissected into two subfamilies (GATA-1,2,3 and GATA-4,5,6), based on their expression levels in different tissues, where only GATA -4, -5 and -6 are associated with cardio- and organogenesis (Pikkarainen et al., 2004; Peterkin et al., 2005; Brewer and Pizzezy, 2006; Whitfield et al., 2012; Turbendian et al., 2013). We found only GATA4 and GATA6 to be expressed. Interactions between GATA-factors and AP-1 are well known, especially co-occurrence of AP-1 together with GATA-4 in several heart cell types and in Leydig cells (Herzig et al., 1997; Suzuki et al., 1999; Schröder et al., 2006; Linnemann et al., 2011; Martin et al., 2012). In our system, GATA6 was expressed in high amounts during the mesoderm induction (< day 3) and early cardiac specification

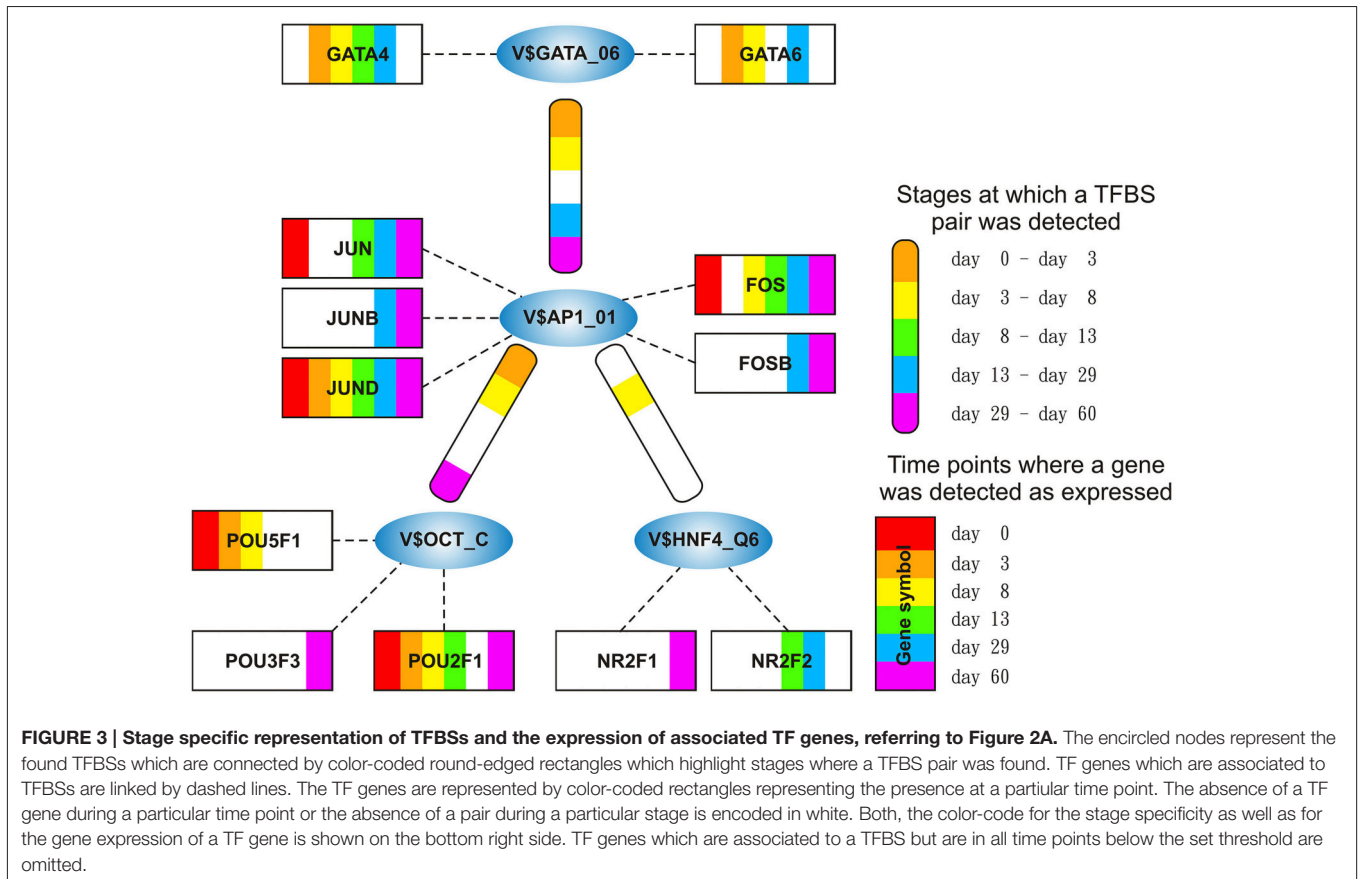


**FIGURE 2 | Clusters we focus on in our analysis in the order in which they are analyzed in this study.** The clusters comprise all interactions during the complete time course, identified by employing MatrixCatch and MCL. The constitution of each cluster for a particular stage is shown in the corresponding tables. **(A)** shows the AP-1-cluster, **Table 1**; **(B)** HMG1Y-cluster, **Table 2**; **(C)** SMAD-cluster, **Table 4**; **(D)** NFAT-cluster, **Table 5**.

**TABLE 1 | TFBS pairs within the AP-1-cluster.**

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V\$AP1_01 – V\$OCT_C	+	+	–	–	+
V\$AP1_01 – V\$GATA_Q6	+	+	–	+	+
V\$AP1_01 – V\$HNF4_Q6	–	+	–	–	–

Constitution of co-occurring pairs in the AP-1-cluster, a “+” indicates the presence of a pair; a “–” its absence. During the late stage of cardiac specification (Day8–Day13), the cluster is completely absent.



stage (day 3–day 8) but was not expressed or only at minor extent during cardiac maturation (> day 13, see **Figure 4C**). In contrast, GATA4 was expressed in high amounts during the late cardiac specification stage as well as during cardiac maturation (> day 8). The missing of AP-1 during mesoderm induction (< day 3) suggests that genes specific for mesoderm induction might be under control of GATA-6, whereas GATA-4 and AP-1 may regulate genes during cardiac maturation (> day 13), synergistically (see Pikkarainen et al., 2004 for the role of GATA-4 and GATA-6).

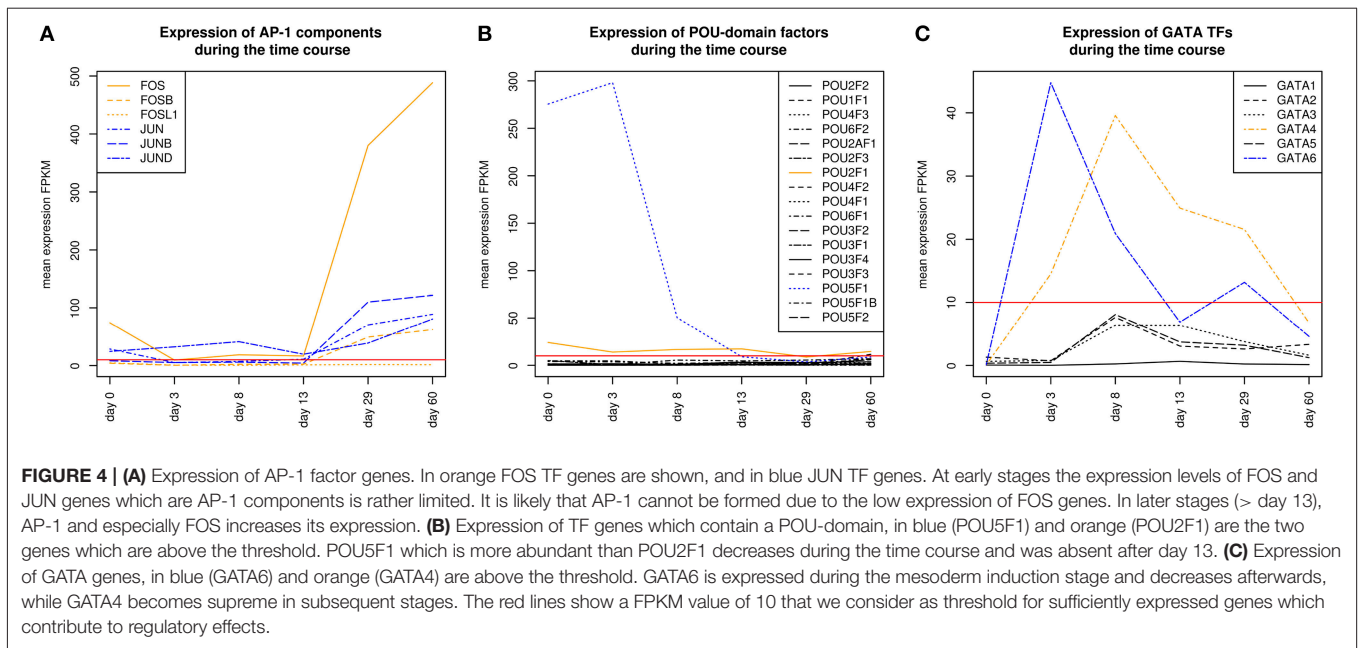
The role of the co-occurrence between V\$AP1\_01 and V\$HNF4\_Q6, which represents a binding site for HNF4A or HNF4G TFs, during cardiomyogenesis is uncertain. This TFBS pair was detected during early cardiac specification stage (days 3–8), but no expression of the related genes could be found. As mentioned before, the formation of AP-1 during this stage at relevant levels is uncertain (see **Figure 4A**), due to the low

expression of the AP-1 components. Furthermore, the role of HNF4-genes, which were frequently reported to be associated with lipid metabolism in the liver (Watt et al., 2003; Chandra et al., 2013), during cardiac development is still unclear, but may point to changes in the metabolism at this stage.

### 3.2. HMGIIY-Cluster

The HMGIIY-cluster is assembled in a total of five TFBS pairs (see **Figures 2B, 5**) with the V\$HMGIIY\_Q6 binding site in its center. **Table 2** shows the co-occurring TFBS pairs of this cluster and **Figure 5** shows for these TFBSs the related TF genes which are expressed in at least one time point. The TFBS pair V\$HMGIIY\_Q6 - V\$OCT\_Q6 was found during all stages and the co-occurrence between V\$HMGIIY\_Q6 and V\$ATF3\_Q6 binding sites was found at days 3–8, and after day 29. Interestingly, we found in this cluster three binding sites, namely V\$NFKAPPAB\_01, V\$NFKB\_Q6\_01, and





**TABLE 2 | TFBS pairs within the HMGIY-cluster.**

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V\$HMGIY_Q6 – V\$OCT_Q6	+	+	+	+	+
V\$HMGIY_Q6 – V\$NFKAPPAB_01	+	+	–	+	+
V\$HMGIY_Q6 – V\$NFKB_Q6_01	+	+	+	+	+
V\$HMGIY_Q6 – V\$NFKB_Q6	+	–	–	–	–
V\$HMGIY_Q6 – V\$ATF3_Q6	+	+	–	–	+

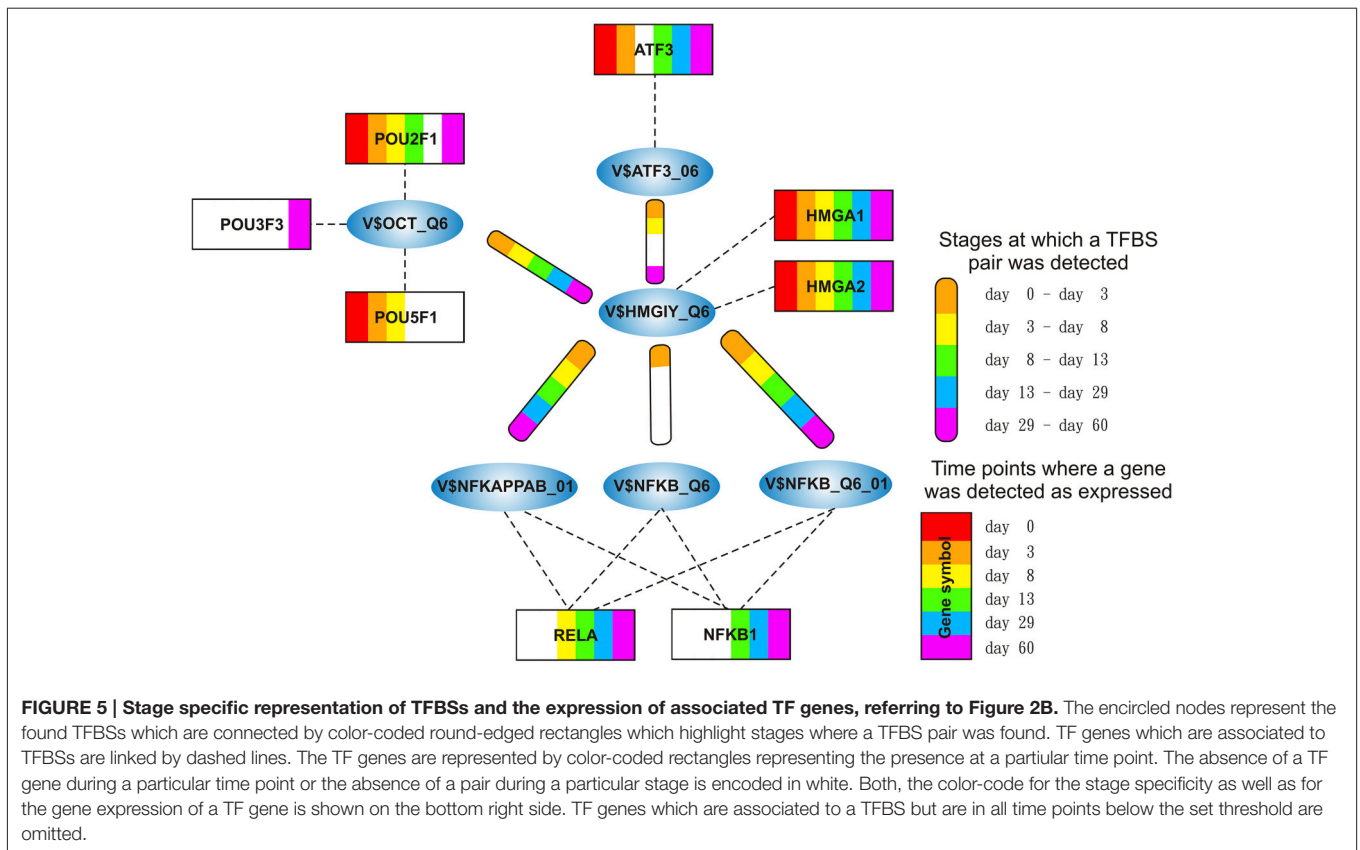
Constitution of co-occurring pairs within the HMGIY-cluster, a “+” indicates the presence of a matrix pair; a “–” its absence.

V\$NFKB\_Q6 which can be bound by the family of NF- $\kappa$ B-related factors. While the V\$HMGIY\_Q6 - V\$NFKB\_Q6 TFBS pair was detected only during the mesoderm induction stage (<day 3), the co-occurrence between V\$HMGIY\_Q6 and V\$NFKB\_Q6\_01 binding sites was found at all stages. The TFBS pair V\$HMGIY\_Q6 - V\$NFKAPPAB\_01 was found at all stages except the late cardiac specification stage (day 8–day 13). To ensure the quality of these three NF- $\kappa$ B binding sites, we further investigated their position weight matrices (PWMs) as well as their binding motifs. Considering the PWMs, we observed that all PWMs have relatively high value of information content (see **Table 3**) which assess their quality. In addition, a comparison between motifs shows different binding behavior of NF- $\kappa$ B-related factors which could be linked to specific members of this family.

HMGA1 is a TF which is represented by the PWM V\$HMGIY\_Q6 and was recently described as a positive regulator of pluripotency in cellular reprogramming (Shah et al., 2012). The expression levels of HMGA1 in our system are in agreement with previous studies, which describe HMGA1 as highly abundant during embryogenesis, especially in embryonic stem cells; with intermediate expression levels in undifferentiated cancers and at low or at not detectable levels in adult

differentiated cells and fibroblasts (Fusco and Fedele, 2007; Hillion et al., 2008, 2009; Resar, 2010; Chou et al., 2011; Schuldenfrei et al., 2011; Shah et al., 2012; Williams et al., 2015). The detected co-occurrence between V\$HMGIY\_Q6 and V\$OCT\_Q6 binding sites was found at all stages. The corresponding TF genes (HMGA1, HMGA2, and POU5F1) of this TFBS pair did not show such behavior (see **Figures 4B, 6A**). HMGA1 as well as POU5F1 are expressed at high levels during early cardiac development with their maximum expression levels at day 3 and declined afterwards. However, this pair was found at later stages indicating that the detected DEGs at these stages could be potentially regulated by this pair. POU5F1 is below the threshold after day 13, whereas HMGA1 is always above the threshold but stabilized at low levels. After day 13, HMGA1, which is in its expression values always more abundant than HMGA2, could regulate the detected pairs alone.

The co-occurrence of V\$HMGIY\_Q6 and different NF- $\kappa$ B binding sites was detected at all time points (see **Table 2**). Interestingly, our findings show that this interaction could occur based on different NF- $\kappa$ B binding sites which are bound by the same TFs. It is known that the interaction between HMGA1 and NF- $\kappa$ B plays a pivotal role in formation of an enhancer complex which is essential to regulate interferon- $\beta$  signaling on



genomic level (Thanos and Maniatis, 1992; Lewis et al., 1994; Wood et al., 1995; Himes et al., 1996; Thanos and Maniatis, 1996; Mantovani et al., 1998; Perrella et al., 1999; Zhang and Verdine, 1999). Within this complex, NF- $\kappa$ B acts on the one hand as a key regulator in hypertrophy and, on the other hand it acts as cardioprotective factor during embryogenesis (Dewey et al., 2011; Gordon et al., 2011; Liu et al., 2012; Zhou et al., 2013). The expression levels of NF- $\kappa$ B genes may indicate an increasing importance of NFKB1 and especially of RELA during cardiac maturation (> day 13), where it is expressed at considerable levels (see Figure 6B).

The co-occurrence of V\$HMG1Y\_Q6 with the V\$ATF3\_Q6 binding site, which is bound by ATF3, was detected during early cardiac development until day 8 and at the latest stage after day 29. ATF-3 is a FOS-related TF, which contains a basic leucine zipper as structural motif (Chen et al., 1994). ATF-3 acts as homo- or heterodimer to activate or to repress the expression of target genes, depending on its environment. Further, it is also involved in TGF- $\beta$  signaling in several cell types and in cardiac development (Ishiguro et al., 2000; Mayr and Montminy, 2001; Yan et al., 2005; Gilchrist et al., 2006; Yin et al., 2010; Lin et al., 2014). While HMGA1 is expressed at high levels during early stages (days 0–3) and is declined afterwards, the ATF3 gene is close to the threshold before day 13 and increases its expression levels during subsequent stages (see Figure 6C). Our results suggest that the genes regulated by this pair are under control of HMGA1 in the early stages and ATF-3 afterwards. Gilchrist et al.

**TABLE 3 | Binding sites for different NF- $\kappa$ B PWMs found in the HMG1Y-cluster.**

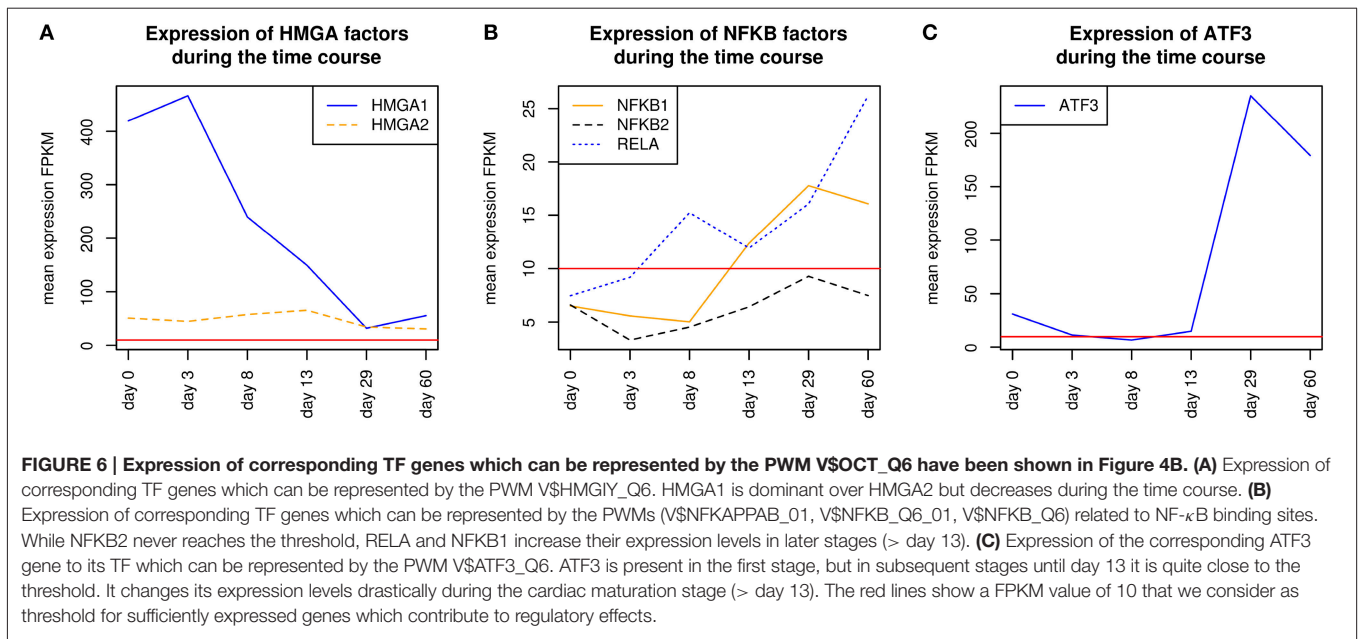
PWM	Information content	Motif
V\$NFKAPPAB_01	11.8	GGGAATTCC
V\$NFKB_Q6_01 <sup>(rc)</sup>	13.3	GGGATTCC
V\$NFKB_Q6	14.4	GGGAAATTCC

The family of NF- $\kappa$ B-related factors can be represented by different PWMs each of which have relatively high information content and different binding motifs. <sup>(rc)</sup>: reverse complement

demonstrate the co-occurrence of ATF-3 and NF- $\kappa$ B binding sites in regulated target genes (Gilchrist et al., 2006). According to their binding sites, our analysis suggests that together with ATF-3 and NF- $\kappa$ B factor, HMGA1 may play an important role in the regulation of target genes in cardiac development.

### 3.3. SMAD-Cluster

The SMAD-cluster is assembled in a total of three TFBS pairs with the V\$SMAD\_Q6\_01 binding site in its center (see Figures 2C, 7). Table 4 shows the co-occurrence of V\$SMAD\_Q6\_01 and V\$FOX\_Q2 binding sites in the promoters of the regulated genes and was observed during all stages. The TFBS pair V\$SMAD\_Q6\_01 - V\$AP1FJ\_Q2 was detected in our system at early stages until day 8 and at late stages after day



13, but not during late cardiac specification stage (days 8–13). In contrast, the co-occurrence between V\$SMAD\_Q6\_01 and V\$LEF1TCF1\_Q4 was detected only during cardiac specification (days 3–13). In addition, **Figure 7** shows for these TFBSs the related TF genes which are expressed in at least one time point.

SMADs are members of a family of transcription factors that form a beta-hairpin structure which interacts with the major groove of the DNA (Burke et al., 1976; Macias et al., 2015). SMAD1-4 which can be represented by the PWM V\$SMAD\_Q6\_01 act as TFs in the nucleus and as signaling molecules, where they are involved in numerous pathways like canonical and non-canonical SMAD-signaling pathways, TGF- $\beta$ - as well as BMP- and WNT-signaling (Heldin et al., 1997; Leask and Abraham, 2004; Euler-Taimor and Heger, 2006; Pal and Khanna, 2006; Schröder et al., 2006; Leask, 2007; Ruiz-Ortega et al., 2007; Calvieri et al., 2012; Massagué, 2012; Dyer et al., 2014; Euler, 2015). **Figure 8A** shows that SMAD1, SMAD2, and SMAD4 genes are continuously expressed at all stages. The detected SMAD3 expression after day 3 exceeds the set threshold only slightly. SMAD2 and SMAD4 show the highest expression levels in our system, but the differences in their expression levels are rather small.

The co-occurrence of V\$SMAD\_Q6\_01 and V\$FOX\_Q2 binding sites was detected at all stages (see **Table 4**). Recently, the cooperative regulatory interaction of FOX factors, which play an important role in cardiovascular development and in other organs (Yamagishi et al., 2003; Maeda et al., 2006; Seo and Kume, 2006; Fortin et al., 2015), with SMAD3 and SMAD4 has been shown by (Fortin et al., 2015). Although the SMAD-FOX pair can be detected during the whole time course, the expression of FOX-genes is limited to FOXH1, which seems to play a role in early heart development only (< day 13, see **Figure 8C**).

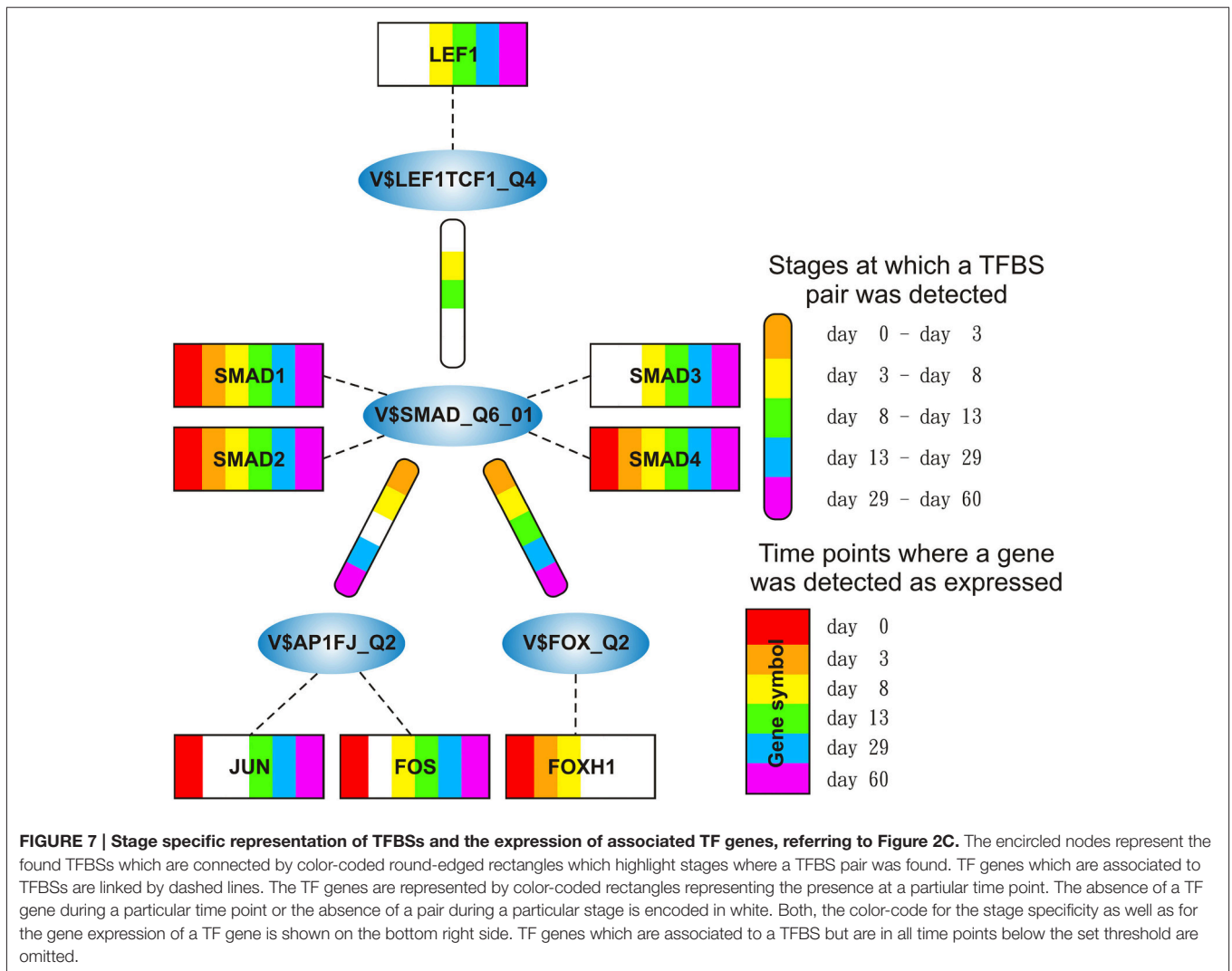
The co-occurrence between V\$SMAD\_Q6\_01 and V\$AP1FJ\_Q2 binding sites were found in almost all stages

except for the late cardiac specification stage (between day 8 and day 13). In adult CMs, AP-1 together with SMAD proteins modulates hypertrophic, apoptotic and fibrotic pathways. Additionally, AP-1 together with SMAD forces the shift toward apoptosis after stimulation of TGF- $\beta$ -signaling (Schneiders et al., 2005; Schröder et al., 2006; Euler, 2015). In the embryonic hearts, the activation of TGF- $\beta$ -pathways results in an induction of cardioprotective functions (Leask and Abraham, 2004; Pal and Khanna, 2006; Leask, 2007; Ruiz-Ortega et al., 2007; Calvieri et al., 2012; Euler, 2015). Although there is no known AP-1 SMAD interaction during cardiogenesis, Yuan et al., shows the interaction of these TFs by usage of AP-1 and SMAD decoy oligodeoxynucleotides, which reduces fibrosis in their study (Yuan et al., 2013).

The detected TFBS pair V\$SMAD\_Q6\_01 - V\$LEF1TCF1\_Q4 is limited to the cardiac specification stage (day 3–day 13). TCF-7 and LEF-1 transcription factors, which are represented by V\$LEF1TCF1\_Q4, can be activated by  $\beta$ -catenin and are involved in canonical WNT-signaling (Brade et al., 2006; Chen et al., 2006; Pal and Khanna, 2006; Kwon et al., 2007; Naito et al., 2010). The measured gene expression of TCF as well as LEF genes shows that during cardiac specification both groups are quite close to or below the set threshold (see **Figure 8B**). This indicates that no TCF or LEF binding occurs, which may result in the absence of canonical WNT-signaling during cardiac specification.

### 3.4. NFAT-Cluster

The NFAT-cluster consists in a total of six TFBS pairs with V\$NFAT\_Q6 binding site in its center (see **Figures 2D, 9**). As described in **Table 5** and **Figure 9**, V\$NFAT\_Q6 co-occurs with V\$PEBP6\_Q6 and V\$ETS1\_B binding sites only during the mesoderm induction stage (days 0–3). Three TFBS pairs, namely V\$NFAT\_Q6 - V\$AP1\_C, V\$NFAT\_Q6 - V\$CREBP1CJUN\_01,



**TABLE 4 | TFBS pairs within the SMAD-cluster.**

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V\$SMAD_Q6_01 – V\$FOX_Q2	+	+	+	+	+
V\$SMAD_Q6_01 – V\$AP1FJ_Q2	+	+	–	+	+
V\$SMAD_Q6_01 – V\$LEF1TCF1_Q4	–	+	+	–	–

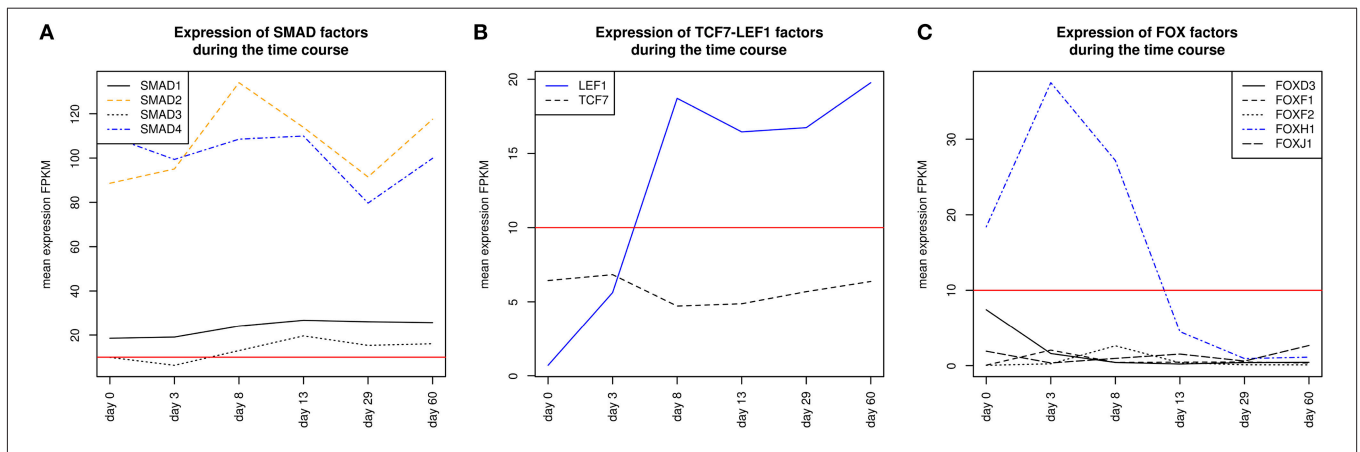
*Constitution of co-occurring pairs within the SMAD-cluster, a “+” indicates the presence of a pair; a “–” its absence.*

and V\$NFAT\_Q6 - V\$MAF\_Q6\_01, were found during the complete time course. The co-occurrence of V\$NFAT\_Q6 with V\$CEBPB\_01 binding sites in the promoter regions of the analyzed set of genes was found as present until day 8 and during the cardiac maturation stage after day 13. This TFBS pair was not present during the late cardiac specification stage (days 8–13). In addition, **Figure 9** shows for these TFBSs the related TF genes which are expressed in at least one time point.

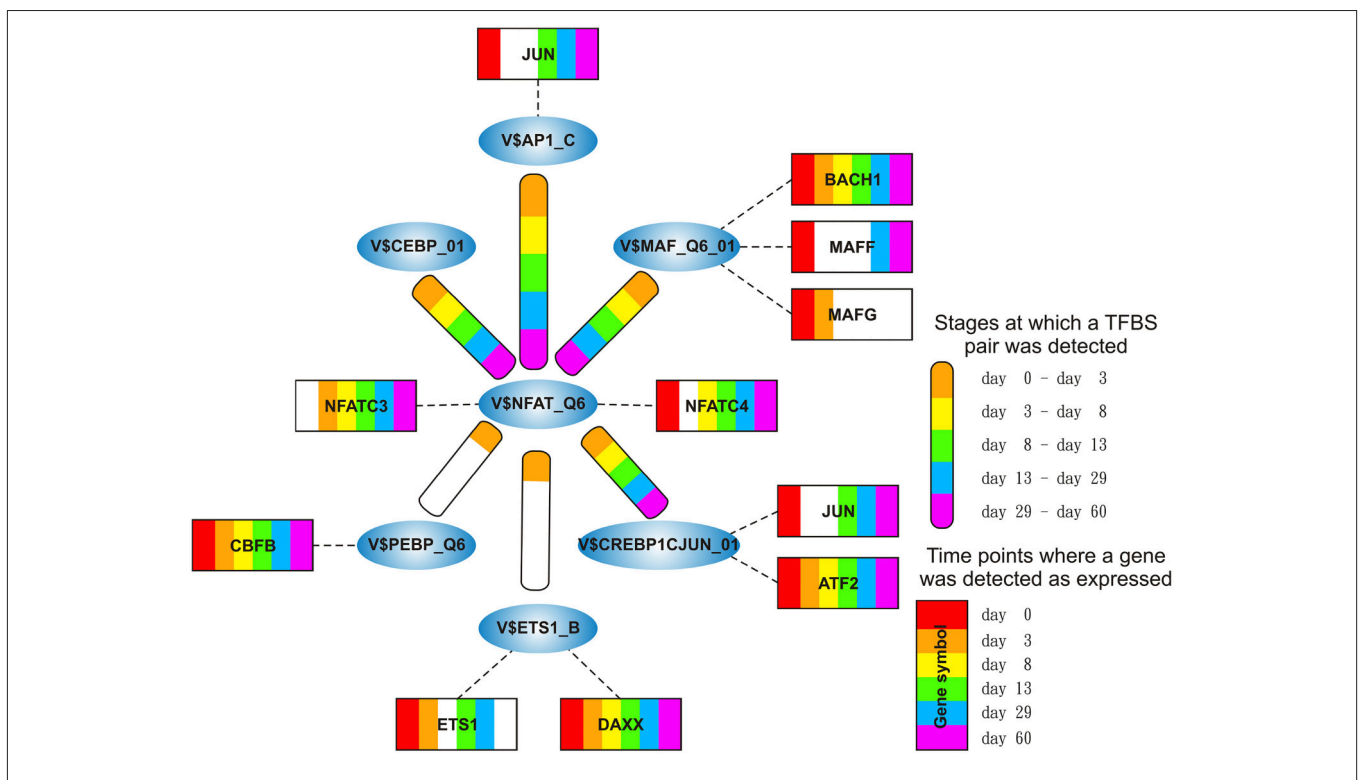
Regulatory roles for NFAT factors, which can be represented by the PWM V\$NFAT\_Q6, have been discovered in diverse organs and cells, including the central nervous system, blood

vessels, heart, skeletal muscle and haematopoietic stem cells (Macián, 2005). In general, an activation of factors of the NFAT family is calcium dependent and has been described to be of specific importance in development of the atrial myocardium and the morphogenesis of heart valves (Graef et al., 2001; Crabtree and Olson, 2002; Schubert et al., 2003; Schulz and Yutzey, 2004). In our system, only NFATC3 and NFATC4 showed expression levels above the threshold. Comparing the expression levels, NFATC4 is more abundant than NFATC3 at all time points, except for day 3, but both genes increase their expression levels at later stages and especially after day 29 (see **Figure 10A**).





**FIGURE 8 | Expression of corresponding TF genes which can be represented by the PWM V\$AP1FJ\_Q2 have been shown in Figure 4A. (A)** Expression of corresponding genes to TFs which can be represented by the PWM V\$SMAD\_Q6\_01. Each SMAD is expressed during the complete time course at similar levels, while the expression levels of SMAD2/4 are higher than the expression levels of SMAD1/3. After beginning of the cardiac specification (> day 3) SMAD4 is slightly more abundant than SMAD2 and remains in this position. **(B)** Expression of corresponding TF genes which can be represented by the PWM V\$LEF1TCF1\_Q4, TCF7 is below the threshold set by us as a limit for robust transcription while LEF1 is clearly transcribed after the mesoderm induction stage (> day 3). The SMAD-TCF/LEF-pair was found during the cardiac specification stage only (day 3–day 8). **(C)** Expression of corresponding TF genes which can be represented by the PWM V\$FOX\_Q2. FOXH1 is the only expressed gene and present until day 13. The red lines show a FPKM value of 10 that we consider as threshold for sufficiently expressed genes which contribute to regulatory effects.



**FIGURE 9 | Stage specific representation of TFBSs and the expression of associated TF genes, referring to Figure 2D.** The encircled nodes represent the found TFBSs which are connected by color-coded round-edged rectangles which highlight stages where a TFBS pair was found. TF genes which are associated to TFBSs are linked by dashed lines. The TF genes are represented by color-coded rectangles representing the presence at a particular time point. The absence of a TF gene during a particular time point or the absence of a pair during a particular stage is encoded in white. Both, the color-code for the stage specificity as well as for the gene expression of a TF gene is shown on the bottom right side. TF genes which are associated to a TFBS but are in all time points below the set threshold are omitted.

**TABLE 5 | TFBS pairs within the NFAT-cluster.**

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V\$NFAT_Q6 – V\$PEBP_Q6	+	–	–	–	–
V\$NFAT_Q6 – V\$AP1_C	+	+	+	+	+
V\$NFAT_Q6 – V\$CEBPB_01	+	+	–	+	+
V\$NFAT_Q6 – V\$CREBP1CJUN_01	+	+	+	+	+
V\$NFAT_Q6 – V\$MAF_Q6_01	+	+	+	+	+
V\$NFAT_Q6 – V\$ETS1_B	+	–	–	–	–

Constitution of the NFAT-cluster, a “+” indicates the presence of a matrix pair; a “–” its absence.

The detected co-occurrence of TFBS pairs V\$NFAT\_Q6 - V\$AP1\_C and V\$NFAT\_Q6 - V\$PEBP\_Q6 refers either to NFAT-AP-1 or to NFAT-RUNX interactions which have been mainly observed in the immune system (Macián, 2005). Macián et al. have demonstrated that the interaction between NFAT and AP-1 can be linked to calcineurin dependent pathways as well as to regulation of MAP kinase pathways (Macián et al., 2001). Additionally, NFAT and AP-1 cooperate in naïve T-cells with RUNX TFs as well as with NF-κB in the promoter of IL-2 during T-cell activation (see **Figures 10C,E**) (Hermann-Kleiter and Baier, 2010). In our system, the low or absent expression of RUNX indicates no relevance for these factors. However, the corresponding binding site can be also occupied by CBFB, which is associated to congenital heart anomalies and is expressed during all time points (Khan et al., 2006).

We found the co-occurring TFBS pair V\$NFAT\_Q6 - V\$MAF\_Q6\_01 at all stages. For the corresponding factors it has been shown by Hogan et al. that NFAT factors and MAF were able to activate IL-4 promoters (Hogan et al., 2003). Of all TFs linked to V\$MAF\_Q6\_01, BACH1 is expressed at all stages and is always more abundant than the other genes shown in **Figure 10B**. This suggests a synergistic interaction in gene regulation between these factors during the complete time course. Furthermore, the interaction between NFAT and MAF factors was observed simultaneously at classical NFAT-AP-1 interaction sites (Hogan et al., 2003).

The co-occurrence between V\$NFAT\_Q6 and V\$CEBPB\_01 binding sites has been described in liver cell lines by Yang and Chow (2003). The corresponding factors to this pair seem to interact in a formation of a composite enhancer complex (Yang and Chow, 2003). In our system, genes that are linked to V\$CEBPB\_01 binding sites are not expressed (see **Figure 10F**). The observation of this pair and its potential role in heart development remains unclear.

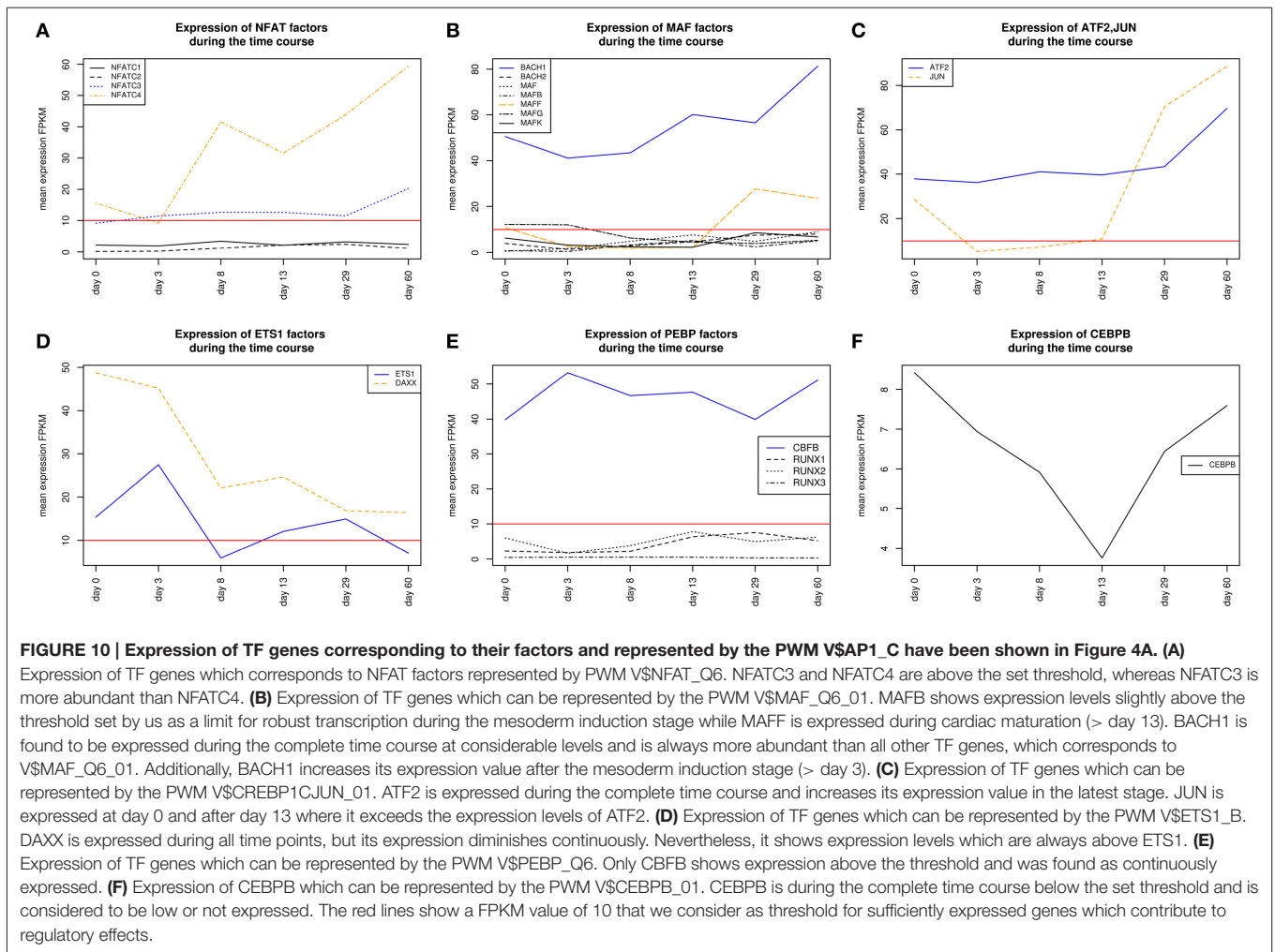
The role of the TFBS pair V\$NFAT\_Q6 - V\$ETS1\_B, which was detected during the mesoderm induction stage, remains unclear. ETS1, a TF gene which can be linked to the PWM V\$ETS1\_B, is required for the differentiation of cardiac neural crest (Gao et al., 2010). Although ETS1 was expressed during the mesoderm induction stage (days 0–3), its expression is markedly reduced afterwards. DAXX is another gene that is linked to the PWM V\$ETS1\_B and is at all time points more abundant than ETS1 (see **Figure 10D**). The DAXX factor inhibits apoptosis in cardiac myocytes (Zobalova et al., 2008). An interaction between NFAT and DAXX was not found in literature, and thus the role of this pair remains unclear.

## 4. DISCUSSION

Today, it is known that in higher organisms transcription factors have to interact with each other to regulate gene expression which leads to a proper development of tissues and organs. So far, several studies have shown that the co-occurrence of TF binding sites (TFBSs) on sequences is an essential indication for the identification of interactions between TFs. In this study, we identified co-occurring TFBS pairs by applying MatrixCatch algorithm to the promoter regions of five differentially expressed gene sets, which are based on a time course dataset of developing human myocardium, modeled in a tissue engineering approach (Hudson et al., in revision). MatrixCatch is a statistically affirmed computational method for the recognition of experimentally verified interactions between TFs according to their TFBS localizations in promoters. However, MatrixCatch recognizes based on its underlying algorithm all detectable TFBS pairs of known interacting TFs in promoter regions. This results in a huge overlap between recognized pairs at different stages, although these pairs can play different roles for each stage. To eliminate this drawback of MatrixCatch to some extent, we created an interaction network based on the TFBS pairs for each stage and then applied the MCL algorithm. MCL differentiates negligible TFBS pairs from densely connected TFBS pairs within these interaction networks and thus determines clusters of TFBSs. Such clusters are important to highlight stage specific co-occurrences of TFBS pairs which provide essential knowledge in the understanding of molecular mechanism of cardiac development.

Additionally, we applied our approach to different lengths of putative promoter regions ([from –500 bp to 0], [from –500 bp to +100 bp], [from –1000 bp to 0]) to determine the influence of promoter lengths on the composition of stage-specific clusters. The results denote that there is a considerably high overlap between stage-specific clusters derived from different putative promoter regions (data not shown). Thus, we considered the –1 kb putative regulatory promoter region for our analysis, which is consistent with our experience and provides the most reliable results.

Although, we filtered MatrixCatch outputs using MCL algorithm to reduce weak co-occurrence of TFBSs in each stage, we detected in our analysis several clusters as well as TFBS pairs whose potential role during cardiac development are unclear. One possible reason for the detection of such pairs could depend on the underlying methodology of MatrixCatch. It uses a computational prediction approach which scans promoter



sequences and their reverse complements to identify TFBSs using PWMs. However, computational identifications of TFBSs generally suffer from high rates of false positive predictions. Another reason for the detection of those clusters or pairs could be due to genes which are expressed at high levels but play different roles in different tissues. As a result, we could identify such clusters or pairs that might play important roles in the regulation of those genes in other tissues but not in heart. For example, we identified the TFBS pair (V\$NFAT\_Q6 - V\$CEBPB\_01) in the NFAT-cluster whose importance has been shown by Yang and Chow in liver (Yang and Chow, 2003), but the potential role of this pair during the cardiac development is unclear. In this context, we also observed the ETS cluster with the V\$ETS\_Q6 binding site in its center (see **Supplementary File 4**). Only some individual components, like ETS factors, in this cluster are associated with potential cardiac functionalities. However, considering TFBS pairs in the ETS cluster, we cannot verify their potential role during the cardiac development.

Our results suggest that different types of co-occurring TFBS pairs can be assigned into two main categories: (i) TFBS pairs which are present in the beginning and in later stages but

absent in at least one of the subsequent stages; (ii) TFBS pairs which are present during all stages. In our clusters presented in the Result section, there are different co-occurring TFBS pairs, like V\$AP1\_01 - V\$OCT\_C and V\$HMGYI\_Q6 - V\$ATF3\_Q6, which fall into the first category. Considering the expression values of TF genes for those pairs, we observed that one TF gene was highly expressed in the beginning stages while its partner is expressed at low levels. After the re-occurrence of such a pair in later stages, the measured expression values of TF genes are exactly the opposite. Consequently, the related TFs cannot act in a synergistic manner but rather in an antagonistic manner. Very drastically, we observed this situation in the expression of AP-1 components and POU5F1, which can be linked to V\$AP1\_01 - V\$OCT\_C TFBS pair (see **Figures 4A,B**). Due to this finding we hypothesize that further TFBS pairs, which fall into the first category, could be helpful to enhance our knowledge on the combinatorial code underlying transcriptional regulation of cardiomyogenesis.

This findings could be discussed in the perspective of the “embryonic hourglass” which describes high divergence in the embryonic shape of vertebrates, insects, like *Drosophila*, and plants, in early and late developmental stages, but minor

divergence in mid-stages (Duboule, 1994; Raff and Wolpert, 1996; Kalinka et al., 2010; Quint et al., 2012). In our study, the number of DEGs as well as the number of identified clusters is high in early stages, converge to a minimum during the late cardiac specification stage (day 8–day 13) and increase afterwards again, which is consistent with the general structure of the hourglass model. Furthermore, the identified TFBS pairs, which fall into the first category, could be separated into two different subsets of genes, the one subset is up-regulated before the late cardiac specification stage, while the other subset is up-regulated afterwards and is supposed to regulated cardiac maturation processes. Our findings support the hourglass model derived by previous findings in *Arabidopsis* as well as several animals (Domazet-Lošo and Tautz, 2010; Kalinka et al., 2010; Quint et al., 2012).

In contrast to the TFBSs pairs in the first category, the co-occurrence of TFBS pairs that fall into the second category seems to indicate a synergistic cooperation between related TFs. In our presented clusters, we obtained several TFBS pairs like V\$HMGY\_Q6 - V\$OCT\_Q6, V\$SMAD\_Q6\_01 - V\$FOX\_Q2, and V\$NFAT\_Q6 - V\$CREBP1CJUN\_01 (for detail see **Tables 2–4**). Considering the expression values of corresponding TF genes for those pairs, we determined that these genes are regulated similarly. For instance, the TF genes HMGA1 and POU5F1, which are linked to V\$HMGY\_Q6 and V\$OCT\_Q6, respectively, are highly expressed during first developmental stages and diminish their levels after day 3. This condition is also observed for the TFBS pair V\$NFAT\_Q6 - V\$CREBP1CJUN\_01 where the associated TF genes are expressed at low levels in the beginning and increase their expression levels in later stages.

Altogether, in our study we performed a systematic analysis of TFBS pairs to address the question of cooperation between TFs linked to TFBS pairs, which could play a crucial role through five different cardiac developmental stages. Addressing this question, our results show that some TFBS pairs can be detected at all developmental stages. Furthermore, we obtained the same TFBS pairs at very early and very late stages of the differentiation, although these stages are completely different in their functions. Especially considering expression values of related TF genes of these pairs, we determined that co-occurrence between TFBSs does not always indicate a synergistic regulation of target genes. This finding suggests that corresponding TFs of these pairs can be bound in a mutual exclusive manner, which is important during cardiac development to differentiate between stem cell programs and later embryonic programs.

## 5. CONCLUSION

We identify transcription factor pairs that drive cardiac development from stem cells to mature cells in a 60 day time course dataset. Our approach is motivated by the importance of potentially interacting transcription factors represented by the co-occurrence of their TFBSs in the regulated stages specific genes and their mediated effects. We identified the relevant pairs employing MatrixCatch method with Markov clustering algorithm together to highlight stage specific clusters of co-occurring TFBS pairs. Furthermore, we analyzed the changes

within these clusters to show the specificity of the gene regulation in cardiac development. Our results demonstrate that similar pairs potentially regulate different developmental stages depending on the expression values of the corresponding genes. This may define switches between embryonic and maturation programs and could contribute to a better understanding of embryonic cardiac development.

## AUTHOR CONTRIBUTIONS

SZ, CM, and MG participated in the design of the study, conducted computational and statistical analyses. EW supervised the computational and statistical analyses. AR, FR prepared the time course data and the experiments. WZ supervised the experimental design and the experiments. SU prepared and processed the RNAseq data which are used in this study. SZ, CM, RT, and MG were involved in interpretation of the results and the literature survey. MG and SZ wrote the final version of the manuscript. MG conceived of and managed the project. All authors read and approved the final manuscript.

## FUNDING

SZ was funded by Mediomics (Fördernummer: 01DJ13026B) of the BMBF (German Ministry of Education and Research) and the DZHK. CM was funded by ExiTox (Fördernummer: 031A269C) of the BMBF (German Ministry of Education and Research). WHZ is supported by the DZHK, the German Research Foundation (DFG ZI 708/10-1, SFB 1002 TP C04/S, and SFB 937 A18), the Foundation Leducq, the German Federal Ministry for Science and Education (BMBF FKZ 13GW0007A [BMBF/CIRM ETIII Award]), and the NIH (U01 HL099997).

## ACKNOWLEDGMENTS

We would like to thank Lena Steins and Martin Haubrock for proofreading the manuscript and providing helpful advices and discussions. This work was supported by the DZHK (German Centre for Cardiovascular Research) and by the BMBF (German Ministry of Education and Research). Furthermore, we acknowledge support by the Open Access Publication Funds of the Göttingen University.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fgene.2016.00033>

**Supplementary File 1 | DEGList.csv includes all detected stage-specific DEGs.**

**Supplementary File 2 | A heatmap of stage-specific DEGs.**

**Supplementary File 3 | TFBS pairs found by MatrixCatch for stage-specific DEGs.**

**Supplementary File 4 | The stage-specific networks after application of MatrixCatch and Markov clustering algorithm.**



## REFERENCES

- Akhurst, R. J. (2012). The paradoxical TGF- $\beta$  vasculopathies. *Nat. Genet.* 44, 838–839. doi: 10.1038/ng.2366
- Bergmann, O., Bhardwaj, R. D., Bernard, S., Zdunek, S., Barnabé-Heider, F., Walsh, S., et al. (2009). Evidence for cardiomyocyte renewal in humans. *Science* 324, 98–102. doi: 10.1126/science.1164680
- Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956. doi: 10.1016/j.cell.2005.08.020
- Brade, T., Männer, J., and Kühl, M. (2006). The role of Wnt signalling in cardiac development and tissue remodelling in the mature heart. *Cardiovasc. Res.* 72, 198–209. doi: 10.1016/j.cardiores.2006.06.025
- Brand, T. (2003). Heart development: molecular insights into cardiac specification and early morphogenesis. *Dev. Biol.* 258, 1–19. doi: 10.1016/S0012-1606(03)00112-X
- Brewer, A., and Pizzev, J. (2006). GATA factors in vertebrate heart development and disease. *Expert. Rev. Mol. Med.* 8, 1–20. doi: 10.1017/S1462399406000093
- Brohée, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinform.* 7:488. doi: 10.1186/1471-2105-7-488
- Buckingham, M., Meilhac, S., and Zaffran, S. (2005). Building the mammalian heart from two sources of myocardial cells. *Nat. Rev. Genet.* 6, 826–835. doi: 10.1038/nrg1710
- Burke, M., Reisler, F., and Harrington, W. F. (1976). Effect of bridging the two essential thiols of myosin on its spectral and actin-binding properties. *Biochemistry* 15, 1923–1927. doi: 10.1021/bi00654a020
- Calvieri, C., Rubattu, S., and Volpe, M. (2012). Molecular mechanisms underlying cardiac antihypertrophic and antifibrotic effects of natriuretic peptides. *J. Mol. Med. (Berl.)* 90, 5–13. doi: 10.1007/s00109-011-0801-z
- Chandra, V., Huang, P., Potluri, N., Wu, D., Kim, Y., and Rastinejad, F. (2013). Multidomain integration in the structure of the HNF-4 $\alpha$  nuclear receptor complex. *Nature* 495, 394–398. doi: 10.1038/nature11966
- Chaudhry, B., Ramsbottom, S., and Henderson, D. J. (2014). Genetics of cardiovascular development. *Prog. Mol. Biol. Transl. Sci.* 124, 19–41. doi: 10.1016/B978-0-12-386930-2.00002-1
- Chen, B. P., Liang, G., Whelan, J., and Hai, T. (1994). ATF3 and ATF3 $\Delta$  zip. Transcriptional repression versus activation by alternatively spliced isoforms. *J. Biol. Chem.* 269, 15819–15826.
- Chen, X., Shevtsov, S. P., Hsich, E., Cui, L., Haq, S., Aronovitz, M., et al. (2006). The  $\beta$ -catenin/T-cell factor/lymphocyte enhancer factor signaling pathway is required for normal and stress-induced cardiac hypertrophy. *Mol. Cell. Biol.* 26, 4462–4473. doi: 10.1128/MCB.02157-05
- Chou, B.-K., Mali, P., Huang, X., Ye, Z., Doney, S. N., Resar, L. M., et al. (2011). Efficient human iPS cell derivation by a non-integrating plasmid from blood cells with unique epigenetic and gene expression signatures. *Cell. Res.* 21, 518–529. doi: 10.1038/cr.2011.12
- Crabtree, G. R., and Olson, E. N. (2002). NFAT signaling: choreographing the social lives of cells. *Cell* 109 (Suppl.), S67–S79. doi: 10.1016/S0092-8674(02)00699-2
- Dewey, F. E., Perez, M. V., Wheeler, M. T., Watt, C., Spin, J., Langfelder, P., et al. (2011). Gene coexpression network topology of cardiac development, hypertrophy, and failure. *Circ. Cardiovasc. Genet.* 4, 26–35. doi: 10.1161/CIRCGENETICS.110.941757
- Deyneko, I. V., Kel, A. E., Kel-Margoulis, O. V., Deineko, E. V., Wingender, E., and Weiss, S. (2013). MatrixCatch—a novel tool for the recognition of composite regulatory elements in promoters. *BMC Bioinform.* 14:241. doi: 10.1186/1471-2105-14-241
- Didié, M., Christalla, P., Rubart, M., Muppala, V., Döker, S., Unsöld, B., et al. (2013). Parthenogenetic stem cells for tissue-engineered heart repair. *J. Clin. Invest.* 123, 1285–1298. doi: 10.1172/JCI66854
- Domazet-Lošo, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468, 815–818. doi: 10.1038/nature09632
- Dongen, S. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, Netherlands.
- Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev. Suppl.* 135–142. Available online at: <http://dev.biologists.org/content/develop/1994/Supplement/135.full.pdf>
- Dyer, L. A., Pi, X., and Patterson, C. (2014). The role of BMPs in endothelial cell function and dysfunction. *Trends Endocrinol. Metab.* 25, 472–480. doi: 10.1016/j.tem.2014.05.003
- Euler, G. (2015). Good and bad sides of TGF $\beta$ -signaling in myocardial infarction. *Front. Physiol.* 6:66. doi: 10.3389/fphys.2015.00066
- Euler-Taimor, G., and Heger, J. (2006). The complex pattern of SMAD signaling in the cardiovascular system. *Cardiovasc. Res.* 69, 15–25. doi: 10.1016/j.cardiores.2005.07.007
- Fortin, J., Ongaro, L., Li, Y., Tran, S., Lamba, P., Wang, Y., et al. (2015). Minireview: Activin signaling in gonadotropes: What does the FOX say to the SMAD? *Mol. Endocrinol.* 29, 963–977. doi: 10.1210/me.2015-1004
- Fusco, A., and Fedele, M. (2007). Roles of HMGA proteins in cancer. *Nat. Rev. Cancer* 7, 899–910. doi: 10.1038/nrc2271
- Gao, Z., Kim, G. H., Mackinnon, A. C., Flagg, A. E., Bassett, B., Earley, J. U., et al. (2010). Ets1 is required for proper migration and differentiation of the cardiac neural crest. *Development* 137, 1543–1551. doi: 10.1242/dev.047696
- Gilchrist, M., Thorsson, V., Li, B., Rust, A. G., Korb, M., Roach, J. C., et al. (2006). Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* 441, 173–178. doi: 10.1038/nature04768
- Gordon, J. W., Shaw, J. A., and Kirshenbaum, L. A. (2011). Multiple facets of NF- $\kappa$ B in the heart: to be or not to NF- $\kappa$ B. *Circ. Res.* 108, 1122–1132. doi: 10.1161/CIRCRESAHA.110.226928
- Graef, I. A., Chen, F., and Crabtree, G. R. (2001). NFAT signaling in vertebrate development. *Curr. Opin. Genet. Dev.* 11, 505–512. doi: 10.1016/S0959-437X(00)00225-2
- Guo, Y., Costa, R., Ramsey, H., Starnes, T., Vance, G., Robertson, K., et al. (2002). The embryonic stem cell transcription factors Oct-4 and FoxD3 interact to regulate endodermal-specific promoter expression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 3663–3667. doi: 10.1073/pnas.062041099
- Heldin, C. H., Miyazono, K., and ten Dijke, P. (1997). TGF- $\beta$  signalling from cell membrane to nucleus through SMAD proteins. *Nature* 390, 465–471. doi: 10.1038/37284
- Hermann-Kleiter, N., and Baier, G. (2010). NFAT pulls the strings during CD4+ T helper cell effector functions. *Blood* 115, 2989–2997. doi: 10.1182/ablood-2009-10-233585
- Herzig, T. C., Jobe, S. M., Aoki, H., Molkentin, J. D., Cowley, A. W. Jr, Izumo, S., et al. (1997). Angiotensin II type1a receptor gene expression in the heart: AP-1 and GATA-4 participate in the response to pressure overload. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7543–7548. doi: 10.1073/pnas.94.14.7543
- Hess, J., Angel, P., and Schorpp-Kistner, M. (2004). AP-1 subunits: quarrel and harmony among siblings. *J. Cell Sci.* 117(Pt 25), 5965–5973. doi: 10.1242/jcs.01589
- Hillion, J., Dhara, S., Sumter, T. F., Mukherjee, M., Di Cello, F., Belton, A., et al. (2008). The high-mobility group A1a/signal transducer and activator of transcription-3 axis: an achilles heel for hematopoietic malignancies? *Cancer Res.* 68, 10121–10127. doi: 10.1158/0008-5472.can-08-2121
- Hillion, J., Wood, L. J., Mukherjee, M., Bhattacharya, R., Di Cello, F., Kowalski, J., et al. (2009). Upregulation of MMP-2 by HMGA1 promotes transformation in undifferentiated, large-cell lung cancer. *Mol. Cancer Res.* 7, 1803–1812. doi: 10.1158/1541-7786.MCR-08-0336
- Himes, S. R., Coles, L. S., Reeves, R., and Shannon, M. F. (1996). High mobility group protein I(Y) is required for function and for c-Rel binding to CD28 response elements within the GM-CSF and IL-2 promoters. *Immunity* 5, 479–489.
- Hogan, P. G., Chen, L., Nardone, J., and Rao, A. (2003). Transcriptional regulation by calcium, calcineurin, and NFAT. *Genes Dev.* 17, 2205–2232. doi: 10.1101/gad.1102703
- Hu, Z., and Gallo, S. M. (2010). Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics* 11:49. doi: 10.1186/1471-2164-11-49
- Ishiguro, T., Nagawa, H., Naito, M., and Tsuruo, T. (2000). Inhibitory effect of ATF3 antisense oligonucleotide on ectopic growth of HT29 human colon cancer cells. *Jpn. J. Cancer Res.* 91, 833–836. doi: 10.1111/j.1349-7006.2000.tb01021.x

- Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. *Cell* 152, 327–339. doi: 10.1016/j.cell.2012.12.009
- Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., et al. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527, 384–388. doi: 10.1038/nature15518
- Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., et al. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468, 811–814. doi: 10.1038/nature09634
- Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32(Suppl. 1), D493–D496. doi: 10.1093/nar/gkh103
- Kel-Margoulis, O. V., Kel, A. E., Reuter, I., Deineko, I. V., and Wingender, E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* 30, 332–334. doi: 10.1093/nar/30.1.332
- Khan, A., Hyde, R. K., Dutra, A., Mohide, P., and Liu, P. (2006). Core binding factor beta (CBFB) haploinsufficiency due to an interstitial deletion at 16q21q22 resulting in delayed cranial ossification, cleft palate, congenital heart anomalies, and feeding difficulties but favorable outcome. *Am. J. Med. Genet. A* 140, 2349–2354. doi: 10.1002/ajmg.a.31479
- Kirby, M. L. (2002). Molecular embryogenesis of the heart. *Pediatr. Dev. Pathol.* 5, 516–543. doi: 10.1007/s10024-002-0004-2
- Kwon, C., Arnold, J., Hsiao, E. C., Taketo, M. M., Conklin, B. R., and Srivastava, D. (2007). Canonical Wnt signaling is a positive regulator of mammalian cardiac progenitors. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10894–10899. doi: 10.1073/pnas.0704044104
- Leask, A. (2007). TGF $\beta$ , cardiac fibroblasts, and the fibrotic response. *Cardiovasc Res.* 74, 207–212. doi: 10.1016/j.cardiores.2006.07.012
- Leask, A., and Abraham, D. J. (2004). TGF- $\beta$  signaling and the fibrotic response. *FASEB J.* 18, 816–827. doi: 10.1096/fj.03-1273rev
- Lewis, H., Kaszubska, W., DeLamar, J. F., and Whelan, J. (1994). Cooperativity between two NF- $\kappa$ B complexes, mediated by high-mobility-group protein I(Y), is essential for cytokine-induced expression of the E-selectin promoter. *Mol. Cell Biol.* 14, 5701–5709. doi: 10.1128/MCB.14.9.5701
- Lin, H., Li, H.-F., Chen, H.-H., Lai, P.-F., Juan, S.-H., Chen, J.-J., et al. (2014). Activating transcription factor 3 protects against pressure-overload heart failure via the autophagy molecule Beclin-1 pathway. *Mol. Pharmacol.* 85, 682–691. doi: 10.1124/mol.113.090092
- Linnemann, A. K., O'Geen, H., Keles, S., Farnham, P. J., and Bresnick, E. H. (2011). Genetic framework for GATA factor function in vascular biology. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13641–13646. doi: 10.1073/pnas.1108440108
- Liu, Q., Chen, Y., Auger-Messier, M., and Molkentin, J. D. (2012). Interaction between NF $\kappa$ B and NFAT coordinates cardiac hypertrophy and pathological remodeling. *Circ. Res.* 110, 1077–1086. doi: 10.1161/CIRCRESAHA.111.260729
- Macián, F. (2005). NFAT proteins: key regulators of T-cell development and function. *Nat. Rev. Immunol.* 5, 472–484. doi: 10.1038/nri1632
- Macián, F., López-Rodríguez, C., and Rao, A. (2001). Partners in transcription: NFAT and AP-1. *Oncogene* 20, 2476–2489. doi: 10.1038/sj.onc.1204386
- Macias, M. J., Martin-Malpartida, P., and Massagué, J. (2015). Structural determinants of Smad function in TGF- $\beta$  signaling. *Trends Biochem. Sci.* 40, 296–308. doi: 10.1016/j.tibs.2015.03.012
- Maeda, J., Yamagishi, H., McAnally, J., Yamagishi, C., and Srivastava, D. (2006). Tbx1 is regulated by forkhead proteins in the secondary heart field. *Dev. Dyn.* 235, 701–710. doi: 10.1002/dvdy.20686
- Mantovani, F., Covaceuszach, S., Rustighi, A., Sgarra, R., Heath, C., Goodwin, G. H., et al. (1998). NF- $\kappa$ B mediated transcriptional activation is enhanced by the architectural factor HMGI-C. *Nucleic Acids Res.* 26, 1433–1439. doi: 10.1093/nar/26.6.1433
- Martin, J., Afouda, B. A., and Hoppler, S. (2010). Wnt/ $\beta$ -catenin signalling regulates cardiomyogenesis via GATA transcription factors. *J. Anat.* 216, 92–107. doi: 10.1111/j.1469-7580.2009.01171.x
- Martin, L. J., Bergeron, F., Viger, R. S., and Tremblay, J. J. (2012). Functional cooperation between GATA factors and cJUN on the star promoter in MA-10 leydig cells. *J. Androl.* 33, 81–87. doi: 10.2164/jandrol.110.012039
- Massagué, J. (2012). TGF $\beta$  signalling in context. *Nat. Rev. Mol. Cell Biol.* 13, 616–630. doi: 10.1038/nrm3434
- Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor CREB. *Nat. Rev. Mol. Cell Biol.* 2, 599–609. doi: 10.1038/35085068
- Meckbach, C., Tacke, R., Hua, X., Waack, S., Wingender, E., and Gültas, M. (2015). PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinform.* 16:400. doi: 10.1186/s12859-015-0827-2
- Miura, G. I., and Yelon, D. (2013). Cardiovascular biology: play it again, Gata4. *Curr. Biol.* 23, R619–R621. doi: 10.1016/j.cub.2013.06.006
- Naito, A. T., Shiojima, I., and Komuro, I. (2010). Wnt signaling and aging-related heart disorders. *Circ. Res.* 107, 1295–1303. doi: 10.1161/CIRCRESAHA.110.223776
- Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyanopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274–1286. doi: 10.1016/j.cell.2012.04.040
- Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., et al. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 95, 379–391. doi: 10.1016/S0092-8674(00)81769-9
- Odom, D. T., Dowell, R. D., Jacobsen, E. S., Nekludova, L., Rolfe, P. A., Danford, T. W., et al. (2006). Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.* 2:2006.0017. doi: 10.1038/msb4100059
- Ohneda, K., and Yamamoto, M. (2002). Roles of hematopoietic transcription factors GATA-1 and GATA-2 in the development of red blood cell lineage. *Acta Haematol.* 108, 237–245. doi: 10.1159/000065660
- Orkin, S. H. (1992). GATA-binding transcription factors in hematopoietic cells. *Blood* 80, 575–581.
- Pal, R., and Khanna, A. (2006). Role of Smad- and Wnt-dependent pathways in embryonic cardiac development. *Stem Cells Dev.* 15, 29–39. doi: 10.1089/scd.2006.15.29
- Perrella, M. A., Pellacani, A., Wiesel, P., Chin, M. T., Foster, L. C., Ibanez, M., et al. (1999). High mobility group-I(Y) protein facilitates nuclear factor- $\kappa$ B binding and transactivation of the inducible nitric-oxide synthase promoter/enhancer. *J. Biol. Chem.* 274, 9045–9052.
- Pesce, M., and Schöler, H. R. (2001). Oct-4: gatekeeper in the beginnings of mammalian development. *Stem Cells* 19, 271–278. doi: 10.1634/stemcells.19-4-271
- Peterkin, T., Gibson, A., Loose, M., and Patient, R. (2005). The roles of GATA-4, -5 and -6 in vertebrate heart development. *Semin Cell Dev. Biol.* 16, 83–94. doi: 10.1016/j.semcdb.2004.10.003
- Pikkarainen, S., Tokola, H., Kerkelä, R., and Ruskoaho, H. (2004). GATA transcription factors in the developing and adult heart. *Cardiovasc Res.* 63, 196–207. doi: 10.1016/j.cardiores.2004.03.025
- Quint, M., Drost, H.-G., Gabel, A., Ullrich, K. K., Bönn, M., and Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. *Nature* 490, 98–101. doi: 10.1038/nature11394
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Raff, R. A., and Wolpert, L. (1996). The shape of life-genes, development and evolution of animal forms. *Genet. Res.* 68:261.
- Resar, L. M. S. (2010). The high mobility group A1 gene: transforming inflammatory signals into cancer? *Cancer Res.* 70, 436–439. doi: 10.1158/0008-5472.can-09-1212
- Ruiz-Ortega, M., Rodríguez-Vita, J., Sanchez-Lopez, E., Carvajal, G., and Egido, J. (2007). TGF- $\beta$  signaling in vascular fibrosis. *Cardiovasc. Res.* 74, 196–206. doi: 10.1016/j.cardiores.2007.02.008
- Ryan, K., and Chin, A. J. (2003). T-box genes and cardiac development. *Birth Defects Res. C Embryo Today* 69, 25–37. doi: 10.1002/bdrc.10001
- Schleich, J.-M., Abdulla, T., Summers, R., and Houyel, L. (2013). An overview of cardiac morphogenesis. *Arch. Cardiovasc. Dis.* 106, 612–623. doi: 10.1016/j.acvd.2013.07.001
- Schneiders, D., Heger, J., Best, P., Piper, H. M., and Taimor, G. (2005). SMAD proteins are involved in apoptosis induction in ventricular cardiomyocytes. *Cardiovasc. Res.* 67, 87–96. doi: 10.1016/j.cardiores.2005.02.021
- Schöler, H. R., Ruppert, S., Suzuki, N., Chowdhury, K., and Gruss, P. (1990). New type of POU domain in germ line-specific protein Oct-4. *Nature* 344, 435–439.

- Schröder, D., Heger, J., Piper, H. M., and Euler, G. (2006). Angiotensin II stimulates apoptosis via TGF- $\beta$ 1 signaling in ventricular cardiomyocytes of rat. *J. Mol. Med. (Berl)*. 84, 975–983. doi: 10.1007/s00109-006-0090-0
- Schubert, W., Yang, X. Y., Yang, T. T. C., Factor, S. M., Lisanti, M. P., Molkentin, J. D., et al. (2003). Requirement of transcription factor NFAT in developing atrial myocardium. *J. Cell. Biol.* 161, 861–874. doi: 10.1083/jcb.2003.01058
- Schuldenfrei, A., Belton, A., Kowalski, J., Talbot, C. C. Jr, Di Cello, F., Poh, W., et al. (2011). HMGA1 drives stem cell, inflammatory pathway, and cell cycle progression genes during lymphoid tumorigenesis. *BMC Genomics* 12:549. doi: 10.1186/1471-2164-12-549
- Schulz, R. A., and Yutzey, K. E. (2004). Calcineurin signaling and NFAT activation in cardiovascular and skeletal muscle development. *Dev. Biol.* 266, 1–16. doi: 10.1016/j.ydbio.2003.10.008
- Seo, S., and Kume, T. (2006). Forkhead transcription factors, Foxc1 and Foxc2, are required for the morphogenesis of the cardiac outflow tract. *Dev. Biol.* 296, 421–436. doi: 10.1016/j.ydbio.2006.06.012
- Shah, S. N., Kerr, C., Cope, L., Zambidis, E., Liu, C., Hillion, J., et al. (2012). HMGA1 reprograms somatic cells into pluripotent stem cells by inducing stem cell transcriptional networks. *PLoS ONE* 7:e48533. doi: 10.1371/journal.pone.0048533
- Shaulian, E. (2010). AP-1–The Jun proteins: Oncogenes or tumor suppressors in disguise? *Cell Signal.* 22, 894–899. doi: 10.1016/j.cellsig.2009.12.008
- Shaulian, E., and Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nat. Cell. Biol.* 4, E131–E136. doi: 10.1038/ncb0502-e131
- Shi, G., and Jin, Y. (2010). Role of Oct4 in maintaining and regaining stem cell pluripotency. *Stem Cell Res. Ther.* 1:39. doi: 10.1186/scrt39
- Shih, Y.-K., and Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics* 28, i473–i479. doi: 10.1093/bioinformatics/bts370
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 1–25. doi: 10.2202/1544-6115.1027
- Soong, P. L., Tiburcy, M., and Zimmermann, W.-H. (2012). Cardiac differentiation of human embryonic stem cells and their assembly into engineered heart muscle. *Curr. Protoc. Cell Biol.* Chapter 23:Unit23.8. doi: 10.1002/0471143030.cb230855
- Suzuki, Y. J., Ikeda, T., Shi, S. S., Kitta, K., Kobayashi, Y. M., Morad, M., et al. (1999). Regulation of GATA-4 and AP-1 in transgenic mice overexpressing cardiac calcineurin. *Cell Calcium* 25, 401–407.
- Sylva, M., van den Hoff, M. J. B., and Moorman, A. F. M. (2014). Development of the human heart. *Am. J. Med. Genet. A* 164A, 1347–1371. doi: 10.1002/ajmg.a.35896
- Takeuchi, T. (2014). Regulation of cardiomyocyte proliferation during development and regeneration. *Dev. Growth. Differ.* 56, 402–409. doi: 10.1111/dgd.12134
- Thanos, D., and Maniatis, T. (1992). The high mobility group protein HMG I(Y) is required for NF- $\kappa$ B-dependent virus induction of the human IFN- $\beta$  gene. *Cell* 71, 777–789.
- Thanos, D., and Maniatis, T. (1996). In vitro assembly of enhancer complexes. *Methods Enzymol* 274, 162–173.
- Tiburcy, M., and Zimmermann, W.-H. (2014). Modeling myocardial growth and hypertrophy in engineered heart muscle. *Trends Cardiovasc. Med.* 24, 7–13. doi: 10.1016/j.tcm.2013.05.003
- Turbendian, H. K., Gordillo, M., Tsai, S.-Y., Lu, J., Kang, G., Liu, T.-C., et al. (2013). GATA factors efficiently direct cardiac fate from embryonic stem cells. *Development* 140, 1639–1644. doi: 10.1242/dev.093260
- Vlasblom, J., and Wodak, S. J. (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 10:99. doi: 10.1186/1471-2105-10-99
- Wang, X., and Jauch, R. (2014). OCT4: A penetrant pluripotency inducer. *Cell Regen (Lond)*. 3:6. doi: 10.1186/2045-9769-3-6
- Watt, A. J., Garrison, W. D., and Duncan, S. A. (2003). HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology* 37, 1249–1253. doi: 10.1053/jhep.2003.50273
- Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., et al. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 13:R50. doi: 10.1186/gb-2012-13-9-r50
- Williams, M. D., Zhang, X., Belton, A. S., Xian, L., Huso, T., Park, J.-J., et al. (2015). HMGA1 drives metabolic reprogramming of intestinal epithelium during hyperproliferation, polyposis, and colorectal carcinogenesis. *J. Proteome Res.* 14, 1420–1431. doi: 10.1021/pr501084s
- Wong, K.-C., Li, Y., and Peng, C. (2016). Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells. *Bioinformatics* 32, 321–324. doi: 10.1093/bioinformatics/btv555
- Wood, L. D., Farmer, A. A., and Richmond, A. (1995). HMG(I)Y and Sp1 in addition to NF- $\kappa$ B regulate transcription of the MGSA/GRO $\alpha$  gene. *Nucleic Acids Res.* 23, 4210–4219. doi: 10.1093/nar/23.20.4210
- Yamagishi, H., Maeda, J., Hu, T., McAnally, J., Conway, S. J., Kume, T., et al. (2003). Tbx1 is regulated by tissue-specific forkhead proteins through a common Sonic hedgehog-responsive enhancer. *Genes Dev.* 17, 269–281. doi: 10.1101/gad.1048903
- Yan, C., Lu, D., Hai, T., and Boyd, D. D. (2005). Activating transcription factor 3, a stress sensor, activates p53 by blocking its ubiquitination. *EMBO J.* 24, 2425–2435. doi: 10.1038/sj.emboj.7600712
- Yang, T. T. C., and Chow, C.-W. (2003). Transcription cooperation by NFAT.C/EBP composite enhancer complex. *J. Biol. Chem.* 278, 15874–15885. doi: 10.1074/jbc.M211560200
- Ye, L., Zimmermann, W.-H., Garry, D. J., and Zhang, J. (2013). Patching the heart: cardiac repair from within and outside. *Circ. Res.* 113, 922–932. doi: 10.1161/CIRCRESAHA.113.300216
- Yin, X., Wolford, C. C., Chang, Y.-S., McConoughey, S. J., Ramsey, S. A., Aderem, A., et al. (2010). ATF3, an adaptive-response gene, enhances TGF $\beta$  signaling and cancer-initiating cell features in breast cancer cells. *J. Cell Sci.* 123(Pt 20), 3558–3565. doi: 10.1242/jcs.064915
- Yuan, H.-F., Huang, H., Li, X.-Y., Guo, W., Xing, W., Sun, Z.-Y., et al. (2013). A dual AP-1 and SMAD decoy ODN suppresses tissue fibrosis and scarring in mice. *J. Invest. Dermatol.* 133, 1080–1087. doi: 10.1038/jid.2012.443
- Zhang, X. M., and Verdine, G. L. (1999). A small region in HMG I(Y) is critical for cooperation with NF- $\kappa$ B on DNA. *J. Biol. Chem.* 274, 20235–20243. doi: 10.1074/jbc.274.29.20235
- Zhou, H., Yang, H.-X., Yuan, Y., Deng, W., Zhang, J.-Y., Bian, Z.-Y., et al. (2013). Paeoniflorin attenuates pressure overload-induced cardiac remodeling via inhibition of TGF $\beta$ /Smads and NF- $\kappa$ B pathways. *J. Mol. Histol.* 44, 357–367. doi: 10.1007/s10735-013-9491-x
- Zimmermann, W.-H., Melnychenko, I., Wasmeier, G., Didić, M., Naito, H., Nixdorff, U., et al. (2006). Engineered heart tissue grafts improve systolic and diastolic function in infarcted rat hearts. *Nat. Med.* 12, 452–458. doi: 10.1038/nm1394
- Zobalova, R., Swettenham, E., Chladova, J., Dong, L.-F., and Neuzil, J. (2008). Daxx inhibits stress-induced apoptosis in cardiac myocytes. *Redox. Rep.* 13, 263–270. doi: 10.1179/135100008X308975

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Zeidler, Meckbach, Tacke, Raad, Roa, Uchida, Zimmermann, Wingender and Gültas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

#### **A.4. A novel sequence-based feature for the identification of DNA-binding sites in proteins using Jensen-Shannon Divergence**



Article

# A Novel Sequence-Based Feature for the Identification of DNA-Binding Sites in Proteins Using Jensen–Shannon Divergence

Truong Khanh Linh Dang <sup>1</sup>, Cornelia Meckbach <sup>2</sup>, Rebecca Tacke <sup>2</sup>, Stephan Waack <sup>1</sup> and Mehmet Gültas <sup>1,\*</sup>

<sup>1</sup> Institute of Computer Science, University of Göttingen, Göttingen 37077, Germany; ldang1@informatik.uni-goettingen.de (T.K.L.D.); waack@informatik.uni-goettingen.de (S.W.)

<sup>2</sup> Institute of Bioinformatics, University Medical Center Göttingen, Göttingen 37077, Germany; c.meckbach@bioinf.med.uni-goettingen.de (C.M.); rebecca.tacke@stud.uni-goettingen.de (R.T.)

\* Correspondence: gueltas@informatik.uni-goettingen.de; Tel.: +49-551-39-172055

Academic Editors: Carlos M. Travieso-González and Jesús B. Alonso-Hernández

Received: 30 July 2016; Accepted: 20 October 2016; Published: 24 October 2016

**Abstract:** The knowledge of protein-DNA interactions is essential to fully understand the molecular activities of life. Many research groups have developed various tools which are either structure- or sequence-based approaches to predict the DNA-binding residues in proteins. The structure-based methods usually achieve good results, but require the knowledge of the 3D structure of protein; while sequence-based methods can be applied to high-throughput of proteins, but require good features. In this study, we present a new information theoretic feature derived from Jensen–Shannon Divergence (JSD) between amino acid distribution of a site and the background distribution of non-binding sites. Our new feature indicates the difference of a certain site from a non-binding site, thus it is informative for detecting binding sites in proteins. We conduct the study with a five-fold cross validation of 263 proteins utilizing the Random Forest classifier. We evaluate the functionality of our new features by combining them with other popular existing features such as position-specific scoring matrix (PSSM), orthogonal binary vector (OBV), and secondary structure (SS). We notice that by adding our features, we can significantly boost the performance of Random Forest classifier, with a clear increment of sensitivity and Matthews correlation coefficient (MCC).

**Keywords:** entropy; Jensen–Shannon divergence; Random Forest; DNA-binding sites

## 1. Introduction

Interactions between proteins and DNA play essential roles for controlling of several biological processes such as transcription, translation, DNA replication, and gene regulation [1–3]. An important step to understand the underlying molecular mechanisms of these interactions is the identification of DNA-binding residues in proteins. These residues can provide a great insight into the protein function which leads to gene expression and could also facilitate the generation of new drugs [4,5].

Until now, several groups have published different studies based on either experimental or computational identification of DNA-binding proteins [1,6–11] as well as residues in these proteins [12–23]. However, the usage of experimental approaches for the determination of binding sites is still challenging since they are often demanding, relatively expensive, and time-consuming. To overcome the difficulty of experimental approaches, it is highly desired to develop fast and reliable computational methods for the prediction of DNA-binding residues. For this purpose, several state-of-the-art prediction methods have been developed for the automated identification of those residues. Such methods can be assigned into two main categories: (i) based on the information observed from structure and sequence in a collective manner; (ii) based on the features derived directly

from the amino acid sequence alone (for more detail see reviews [24] and [25]). Although the first type of approaches provides promising information about DNA-binding residues in proteins, their application is difficult due to the limited number of experimentally determined protein structures. In contrast to structure-based approaches, sequence-based methods have been developed by extracting different sequence information features, like amino acid frequency, position-specific scoring matrix (PSSM), BLOSUM62 matrix, sequence conservation, etc. [3,4,18,19,26,27]. Using these features, several machine learning techniques have been applied to construct the classifiers for the prediction of binding residues in proteins. To this end, a variety of support vector machine (SVM) classifiers have been developed in recent studies [2,17–19,23,26,28]. For example, Westhof et al. have recently used an SVM classifier approach in their study, named RBscore (<http://ahsoka.u-strasbg.fr/rbscore/>), by using the physicochemical and evolutionary features that are linearly combined with a residue neighboring network [2]. Further, SVM algorithms were also applied for the models proposed in BindN [18], DISIS [19], BindN+ [23], DP-Bind [27] using different sequence information features including the biochemical property of amino acids, sequence conservation, evolutionary information in terms of PSSM, the side chain pKa value, hydrophobicity index, molecular mass and BLOSUM62 matrix. In addition, other machine learning classifiers such as neural network models [13,15], naive Bayes classifier [26], Random Forest classifiers (RF) [4,29,30] have been developed based on the features derived from protein sequences. For example, Wong et al. [29] have recently developed a successful method using RF classifier with both DNA and protein derived features to predict the specific residue-nucleotide interactions for different DNA-binding domain families.

Despite the rich literature on the sequence-based methods as mentioned above, to date there is still a need to find suitable feature extraction approaches that can enhance the characteristics of DNA-binding residues and thus help to improve the performance of existing methods for identification of DNA-binding residues in proteins. For this aim, we introduce and evaluate a new information theory-based method for the prediction of these residues using Jensen–Shannon divergence (JSD). As a divergence measure based on the Shannon entropy, JSD is a symmetrized and smoothed version of the Kullback–Leibler divergence and is often used for different problems in the field of bioinformatics [31–35]. In this study, following the line of Capra et al. [34] we first quantify the divergence between the observed amino acid distribution of a site in a protein and the background distribution of non-binding sites by using JSD. After that, in analogy to our previous studies QCMF [32] and CMF [36], we incorporate biochemical signals of binding residues in the calculation of JSD that results in the intensification of the DNA-binding residue signals from the non-binding signals.

To demonstrate the performance and functionality of our proposed approach, we apply Random Forest (RF) classifier using our new JSD based features together with three widely used machine learning features, namely position-specific scoring matrix (PSSM), secondary structure (SS) information, and orthogonal binary vector (OBV) information (see review [24]). Our results show that using JSD based features, RF classifier reaches an improved performance in identifying DNA-binding residues with a significantly higher Matthews correlation coefficient (MCC) value in comparison to using previous features alone. Although we only applied RF classifier in this study, both of our sequence-based features could be used in other classifiers such as SVM, neural networks, or decision trees.

## 2. Results

In this study, we introduce new sequence-based features using JSD to improve the performance of previous machine learning approaches in identification of DNA-binding residues in proteins. For this purpose, we propose new sequence-based features ( $f_{\text{JSD}}$  and  $f_{\text{JSD-t}}$ ) using JSD in two different ways. First, using JSD, we calculate the divergences between observed amino acid distributions in multiple sequence alignments (MSAs) of proteins under study and the background distribution which is calculated according to amino acid counts at non-binding residue positions in MSAs. In the second step, we transform the observed amino acid distributions with a doubly stochastic matrix (DSM) to

enhance the weak signal of binding sites in proteins which could not be predicted in the first step. Finally, we calculate for each residue in proteins JSD-based scores and use them for the improvement of the performance of machine learning approaches.

To evaluate our new features, we use two frequently considered cut-off distances of 3.5 Å and 5 Å and thus define a residue in a protein as DNA-binding if the distance between at least one atom on its backbone or side chain and the DNA molecule is smaller than the considered cut-off.

The Results section of this study comprises of two parts. First, we investigate the functionality of our new features combining them in Random Forest (RF) classifier with three previous features. The RF classifier is constructed from 4298 positive and 44,805 negative instances extracted from 263 proteins. The performance of the classifier is evaluated using a five-fold cross validation procedure in which we randomly divided the samples into five parts. The assessment is performed by choosing each of these parts as a test set and the remaining four parts as a training set for model selection. Second, to illustrate the usefulness of our new approach for the prediction of DNA-binding residues, we analyzed the proto-oncogenic transcription factor MYC-MAX (PDB-ID: 1NKP) which is a heterodimer protein complex of two proteins. It is important to note that this protein complex is not included in the training dataset.

### 2.1. Random Forest Classifier

To apply the Random Forest (RF) classifier, we combine our new features ( $f_{\text{JSD}}$  and  $f_{\text{JSD-t}}$ ) with the features  $f_{\text{PSSM}}$ ,  $f_{\text{OBV}}$ , and  $f_{\text{SS}}$  which are widely used for the prediction of DNA-binding residues. Our results show that using our features RF classifier reaches an improved performance in identifying DNA-binding sites with clearly higher statistical values (see Tables 1 and 2). Moreover, we individually evaluated the combination of our features with existing features. The results suggest that the classifier with  $f_{\text{JSD-t}}$  feature has provided better sensitivity and comparable Matthews correlation coefficient (MCC) values in comparison to  $f_{\text{JSD}}$  feature. However, its specificity is moderately decreased. A further comparison reveals that the usage of our both features together with other features does not affect the performance of the classifier. The details are presented for 3.5 Å in Table 1 and for 5 Å in Table 2 and in Appendix A with the standard error of each of the performance measures over the values obtained in the five iterations (see Tables A1 and A2).

**Table 1.** Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å. The prediction system was evaluated by five-fold cross validation.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
$f_{\text{PSSM}}$	0.292	0.963	0.307	0.777	0.313
$f_{\text{PSSM}} + f_{\text{JSD}}$	0.385	0.949	0.349	0.795	0.369
$f_{\text{PSSM}} + f_{\text{JSD-t}}$	0.41	0.939	0.35	0.802	0.377
$f_{\text{PSSM}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.414	0.94	0.348	0.800	0.376
$f_{\text{PSSM}} + f_{\text{SS}}$	0.339	0.958	0.334	0.794	0.338
$f_{\text{PSSM}} + f_{\text{SS}} + f_{\text{JSD}}$	0.416	0.95	0.378	0.808	0.390
$f_{\text{PSSM}} + f_{\text{SS}} + f_{\text{JSD-t}}$	0.441	0.94	0.372	0.817	0.401
$f_{\text{PSSM}} + f_{\text{SS}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.439	0.94	0.37	0.814	0.399
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}}$	0.367	0.968	0.398	0.838	0.413
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}}$	0.422	0.958	0.409	0.837	0.425
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD-t}}$	0.447	0.95	0.403	0.841	0.431
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.444	0.947	0.393	0.835	0.423

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

**Table 2.** Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å. The prediction system was evaluated by five-fold cross validation.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
$f_{\text{PSSM}}$	0.286	0.966	0.350	0.778	0.425
$f_{\text{PSSM}} + f_{\text{JSD}}$	0.395	0.95	0.407	0.801	0.487
$f_{\text{PSSM}} + f_{\text{JSD-t}}$	0.418	0.943	0.411	0.807	0.494
$f_{\text{PSSM}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.426	0.942	0.414	0.807	0.497
$f_{\text{PSSM}} + f_{\text{SS}}$	0.334	0.963	0.386	0.796	0.455
$f_{\text{PSSM}} + f_{\text{SS}} + f_{\text{JSD}}$	0.424	0.951	0.436	0.814	0.513
$f_{\text{PSSM}} + f_{\text{SS}} + f_{\text{JSD-t}}$	0.448	0.944	0.438	0.820	0.520
$f_{\text{PSSM}} + f_{\text{SS}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.445	0.944	0.434	0.819	0.521
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}}$	0.337	0.975	0.431	0.830	0.517
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}}$	0.419	0.958	0.450	0.832	0.535
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD-t}}$	0.439	0.952	0.453	0.836	0.539
$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.442	0.949	0.445	0.832	0.535

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

To further investigate the performance of JSD-based features proposed in this study, we analyzed two additional datasets, namely RBscore [2] and PreDNA datasets [37]. Although the RBscore and PreDNA datasets initially contain 381 and 224 DNA-binding proteins, respectively, we have eliminated a few proteins since they are either included in our training dataset or ineligible due to their MSAs. Consequently, we constructed RF classifier using 263 proteins (which were also used for cross-validation) and randomly selecting 60 proteins from each dataset for testing, respectively. The results of these analyses consistently suggest that our new features show great complementary effect to the previous features which often leads to clear improvement of the classification performance (see Tables 3 and 4). The detailed performance of classifier on different features using different cut-offs for each dataset can be found in Appendix A (see Tables A3–A6).

Considering the AUC-ROC and AUC-PR as the only evaluation factor, results indicate that the RF classifier often achieved its best performance based on both cut-off distances if we combine our new  $f_{\text{JSD-t}}$  feature together with the existing three features (see Tables 1–3). Interestingly, by analyzing the PreDNA dataset we observed that RF classifier with  $f_{\text{JSD}}$  or  $f_{\text{JSD-t}}$  features for the cut-off of 3.5 Å showed similar performance. However, regarding to the distance cut-off of 5 Å, the classifier with  $f_{\text{JSD}}$  feature reached slightly better performance than those with  $f_{\text{JSD-t}}$  feature (see Table 4). After looking at the overall performances, it is inferred that adding our new features can boost the performance of the RF classifier in terms of AUC-ROC and AUC-PR.

**Table 3.** Prediction performance of Random Forest (RF) classifier on RBscore dataset using different distance cut-offs.

Cut-Off	Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
3.5 Å	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}}$	0.517	0.976	0.534	0.896	0.528
	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}}$	0.58	0.967	0.54	0.907	0.543
	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD-t}}$	0.612	0.963	0.546	0.910	0.551
	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.601	0.962	0.531	0.909	0.546
5.0 Å	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}}$	0.499	0.98	0.584	0.895	0.641
	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}}$	0.57	0.968	0.595	0.908	0.661
	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD-t}}$	0.592	0.965	0.60	0.908	0.665
	$f_{\text{PSSM}} + f_{\text{OBV}} + f_{\text{SS}} + f_{\text{JSD}} + f_{\text{JSD-t}}$	0.594	0.964	0.597	0.907	0.663

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

**Table 4.** Prediction performance of RF classifier on PreDNA dataset using different distance cut-offs.

Cut-Off	Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
3.5 Å	$f_{PSSM} + f_{OBV} + f_{SS}$	0.428	0.977	0.458	0.867	0.451
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD}$	0.511	0.97	0.488	0.885	0.488
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD-t}$	0.539	0.962	0.475	0.888	0.488
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD} + f_{JSD-t}$	0.539	0.961	0.47	0.886	0.488
5.0 Å	$f_{PSSM} + f_{OBV} + f_{SS}$	0.395	0.98	0.488	0.858	0.530
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD}$	0.48	0.968	0.511	0.874	0.563
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD-t}$	0.506	0.962	0.51	0.873	0.560
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD} + f_{JSD-t}$	0.499	0.96	0.498	0.871	0.555

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

## 2.2. Position Analysis of the MYC-MAX Protein

The proto-oncogenic transcription factor MYC-MAX (PDB-Entry 1NKP) is a heterodimer protein complex that is active in cell proliferation and is over-expressed in many different cancer types [38]. MYC-MAX transcription factors bind to Enhancer boxes (a core element of the promoter that consists of six nucleotides) and activate transcription of the underlying genes [39].

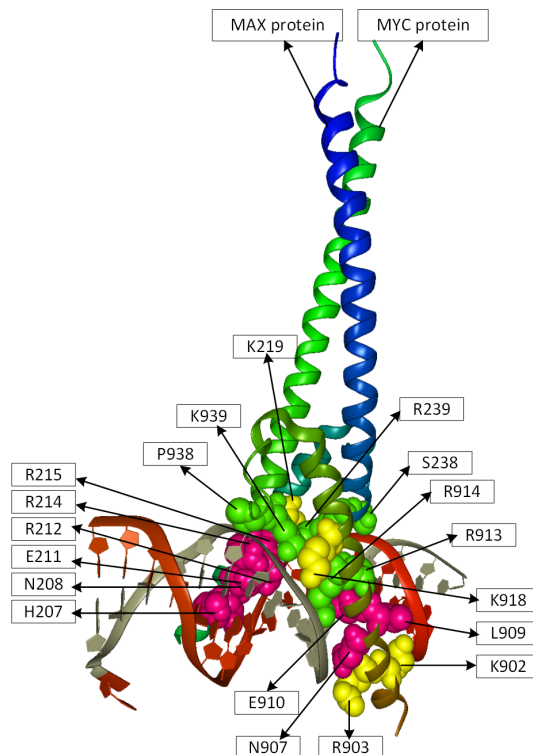
The amino acid chain of MYC protein consists of 88 residues, ten of which are known DNA-binding sites indicating that their distances to DNA are less than 3.5 Å. Applying RF classifier, which takes a majority vote among the random tree classifiers, with our first feature ( $f_{JSD}$ ) combined with existing features, we predicted in total 17 residue positions to be DNA-binding in MYC protein. Seven out of these positions (H906, N907, E910, R913, R914, P938, K939) correspond to the true DNA-binding sites of this protein. While the sites R913, R914, P938, and K939 could also be identified by RF classifier without using our new JSD-based features, the remaining three binding sites could only be detected using our features (for details see Table 5 and Figure 1). Interestingly, using  $f_{JSD-t}$  together with  $f_{PSSM}$ ,  $f_{OBV}$ , and  $f_{SS}$ , the RF classifier correctly predicted these seven positions again as binding sites.

The second protein in the proto-oncogenic transcription factor complex is the MAX protein which consists of 83 residues including nine DNA-binding sites. Using  $f_{JSD}$  or  $f_{JSD-t}$  together with existing features individually, we observed 14 and 13 residue positions to be DNA-binding in MAX protein, respectively. Eight of the predicted positions (H207, N208, E211, R212, R214, R215, S238, R239) found by using either of our both features are true DNA-binding sites in MAX protein. However, without using our new features the RF classifier could only identify two (S238, R239) out of nine true DNA-binding sites in MAX protein (for details see Table 5 and Figure 1). Further, we observed that, the usage of  $f_{JSD-t}$  leads to the reduction of false positive predictions in identifying DNA-binding sites in MAX protein.

**Table 5.** Prediction performance of RF classifier on different features using a cut-off of 3.5 Å for MYC-MAX protein complex (Protein Data Bank (PDB)-Entry 1NKP).

Protein	Feature	Sensitivity	Specificity	MCC
MYC	$f_{PSSM} + f_{OBV} + f_{SS}$	0.30	0.941	0.282
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD}$	0.70	0.853	0.448
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD-t}$	0.70	0.853	0.448
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD} + f_{JSD-t}$	0.70	0.868	0.470
MAX	$f_{PSSM} + f_{OBV} + f_{SS}$	0.222	1.0	0.447
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD}$	0.888	0.906	0.664
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD-t}$	0.888	0.922	0.697
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{JSD} + f_{JSD-t}$	0.889	0.922	0.697

MCC: Matthews correlation coefficient.



**Figure 1.** DNA-binding sites in proto-oncogenic transcription factor MYC-MAX protein complex (PDB-Entry 1NKP). Green spheres denote positions of the DNA-binding sites in both proteins which are detected by RF classifier either using the existing features ( $f_{PSSM}$ ,  $f_{OBV}$ , and  $f_{SS}$ ) alone or combining our new features with these existing features together. Purple spheres show the localization of additional binding sites which were only found by RF classifier using our new features with existing features. Moreover, there are further three binding sites in MYC protein and one binding site in MAX protein, shown with yellow spheres, that could not be identified by the classifier.

Moreover, when statistically evaluating both of our features, we observed that using our sequence-based features RF classifier reaches a significantly improved performance in identifying DNA-binding sites of both proteins with significantly higher sensitivity and MCC values whereas the specificity is moderately decreased. The simultaneous usage of both of our features together with  $f_{PSSM}$ ,  $f_{OBV}$ , and  $f_{SS}$  could result in the decrement of specificity or MCC values. The details are presented in Table 5.

### 3. Materials and Methods

In this section, we describe in particular the data we have used and our new residue-wise features designed to predict DNA-binding sites in proteins.

#### 3.1. Materials

To compile our data needed for training and test, we started with the DBP-374 data set of representative protein-DNA complexes from the Protein Data Bank (PDB) [40] published by Wu et al. [5]. Having performed a comparison with the new PDB version, we calculate for every remaining protein a multiple sequence alignment (MSA) using HHblits and the UniProt20 database (version from June 2015) [41]. We eliminated all proteins, the MSA of which has less than 125 rows, so that we finally ended up with a dataset of 263 protein-DNA complexes and associated MSAs. To obtain our results we perform a five-fold cross validation.



As in [5], an amino acid residue is regarded as a binding site, if it contains at least one atom at distance of less than or equal to 3.5 Å or 5 Å from any atom of DNA molecule in the DNA-protein complex. Otherwise it is treated as non-binding site. For the distance cut-off of 3.5 Å, our set contains 4298 binding sites and 44,805 non-binding sites. For the distance cut-off of 5 Å, however, our data set contains 7211 binding sites and 41,892 non-binding sites.

### 3.2. Methods

Let  $M$  be a multiple sequence alignment, where its first row represents the protein under study. Every residue of that protein is then uniquely determined by its column. In what follows, we identify the residues of the protein with their columns of the MSA.

Grosse et al. [35] pointed out that the Jensen–Shannon divergence (JSD) is extremely useful when it comes to discriminate between two (or more) sources. Capra and Singh [34] carefully discussed several information theoretic measures like Shannon entropy, von Neumann entropy, relative entropy, and sum-of-pair measures to assess sequence conservation. They were the first using JSD in this context and stated its superiority. Gültas et al. [32] showed that the Jensen–Shannon divergence in the context of quantum information theory is of remarkable power. These three articles encouraged us to use JSD in this study. Our first idea is to design a new feature for the prediction of DNA-binding sites in proteins which leverages the *Jensen–Shannon divergence*

$$\text{JSD}(\mathbf{p}_k \parallel \mathbf{p}_{nd}) := \mathbb{H}((\mathbf{p}_k + \mathbf{p}_{nd})/2) - (\mathbb{H}(\mathbf{p}_k) + \mathbb{H}(\mathbf{p}_{nd}))/2. \quad (1)$$

Therein,  $\mathbf{p}_k$  is the empirical amino acid distribution of the  $k$ -th column of the query MSA  $M$ , and  $\mathbf{p}_{nd}$  is the *null distribution* taken over all non-binding sites of our training data.

More precisely, we represent every column  $k$  of every MSA  $M$  considered by a  $20 \times 20$  counting matrix  $C(M_{\cdot,k})$ . The matrix  $C$  is symmetric and its rows as well as columns are indexed by the 20 amino acids. For every ordered pair of amino acids  $(a, a')$ , the matrix coefficient  $C(M_{\cdot,k})_{aa'}$  is equal to the number of ordered pairs  $(i, j)$  ( $i \neq j$ ) of row indices of  $M$  such that  $M_{ik} = a$  and  $M_{jk} = a'$ .

To compute the null distribution  $\mathbf{p}_{nd}$ , we first set up the  $20 \times 20$  counting matrix  $\mathcal{C}_{nd}$  using our training data.  $\mathcal{C}_{nd}$  is the sum over all matrices  $C(M_{\cdot,k})$ , where  $M$  ranges over all training MSAs and  $k$  ranges over all non-binding site columns of  $M$ . Next, the rows of  $\mathcal{C}_{nd}$  are added up. Finally, the resulting row vector is normalized to obtain  $\mathbf{p}_{nd}$ .

There is nothing wrong with the idea that a large value  $\text{JSD}(\mathbf{p}_k \parallel \mathbf{p}_{nd})$  indicates that  $k$  is a DNA-binding residue. However, no information on binding sites is integrated. Only the non-binding sites of our training data are used to compute  $\mathbf{p}_{nd}$ . As we have seen in [32] and [36], transforming empirical amino acid distributions of MSA columns by a carefully designed doubly stochastic matrix is an effective way to integrate the binding site signals. To this end, we first set up a counting matrix  $\mathcal{C}_{bind}$  in a way similar to that of calculating the matrix  $\mathcal{C}_{nd}$ . The difference is that the variable column index  $k$  now ranges over all binding site columns of the training MSAs. Taking the counting matrix  $\mathcal{C}_{bind}$  as input, the doubly stochastic matrix  $\mathcal{D}$  is computed by means of the canonical row-column normalization procedure [42].

Let  $M$  be the query MSA having  $\ell$  columns. Compared with [32] and [36], we enhance the effect of transforming  $M$ 's empirical column distributions by means of the doubly stochastic matrix  $\mathcal{D}$  just defined. Let  $k$  be a column index of  $M$ . First, we compute the matrix product  $C^{(t)}(M_{\cdot,k}) := C(M_{\cdot,k}) \cdot \mathcal{D}$ . Second, we add up all of  $C^{(t)}(M_{\cdot,k})$ 's rows. Finally, we normalize the resulting row to obtain the transformed empirical row distribution  $\mathbf{p}_k^{(t)}$ .

We define two *window scores*  $\text{score}_{\text{JSD},M}(k)$  and  $\text{score}_{\text{JSD-t},M}(k)$  of residue  $k$  w.r.t. query MSA  $M$ , where the window  $\mathfrak{w}(k)$  surrounding  $k$  formally equals  $\{k-3, k-2, k-1, k, k+1, k+2, k+3\} \cap \{1, 2, \dots, \ell\}$ . Clearly, if  $k \in \{4, 5, \dots, \ell-3\}$ ,  $|\mathfrak{w}(k)| = 7$ . Otherwise  $|\mathfrak{w}(k)| \in \{4, 5, 6\}$ . Recapitulate that for any real  $x$  the binomial coefficient  $\binom{x}{2}$  equals  $x(x-1)/2$ . We define the scores as follows.

$$\text{score}_{\text{JSD},M}(k) := \frac{\sum_{l \in \mathfrak{w}(k)} (4 - |k - l|) \text{JSD}(\mathbf{p}_{k+l} \parallel \mathbf{p}_{nd})}{16 - \binom{8 - |\mathfrak{w}(k)|}{2}} \quad (2)$$

$$\text{score}_{\text{JSD-t},M}(k) := \frac{\sum_{l \in \mathfrak{w}(k)} (4 - |k - l|) \text{JSD}(\mathbf{p}_{k+l}^{(t)} \parallel \mathbf{p}_{nd})}{16 - \binom{8 - |\mathfrak{w}(k)|}{2}} \quad (3)$$

The preceding two score definitions are motivated as follows. Bartlett et al. [43] and Panchenko et al. [44] pointed out that exploiting conservation properties of spatial neighbors is useful to predict a residue as functionally important. Since the 3D structures are often unavailable, Capra and Singh [34] developed a window score for such predictions. The concrete shape of our scores takes pattern form Janda et al. [45], who in turn refer to Fischer et al. [33]. Our scores are convex combinations of the Jensen–Shannon terms associated with the residues belonging to the surrounding window  $\mathfrak{w}(k)$ . The weights fall linearly in the distance from  $k$ .

In a last step, we transform two window scores according to Equations (2) and (3) with respect to the query MSA  $M$  into final scores using the Equations (4) and (5), respectively. To this end, for every column index  $k \in \{1, 2, \dots, \ell\}$  of  $M$  we define:

$$f_{\text{JSD},M}(k) := \frac{|\{k' \mid 1 \leq k' \leq \ell, \text{score}_{\text{JSD},M}(k) \geq \text{score}_{\text{JSD},M}(k')\}|}{\ell} \quad (4)$$

$$f_{\text{JSD-t},M}(k) := \frac{|\{k' \mid 1 \leq k' \leq \ell, \text{score}_{\text{JSD-t},M}(k) \geq \text{score}_{\text{JSD-t},M}(k')\}|}{\ell}. \quad (5)$$

The Equations (4) and (5) are basically the determination of the percentage of scores below the current one at index  $k$ . This transformation procedure is essential because it converts MSA-dependent window scores to MSA-independent scores.

To demonstrate the benefit of our new features, we adopt the features  $f_{\text{PSSM}}$ ,  $f_{\text{OBV}}$  and  $f_{\text{SS}}$  devised in [5]. Together with our two new features  $f_{\text{JSD}}$  and  $f_{\text{JSD-t}}$ , we plugged them into the Random Forest (RF) classifier [46] (see Tables 1 and 2 for the combinations we used). For the RF implementation we used the WEKA data mining software [47].

To deal with the imbalanced data problem, we applied bagging techniques suggested in [48]. Since we make use of five-fold cross validation, we randomly split the dataset into 5 roughly equal-sized parts. Every training phase performed on 4 parts consists of 11 sub-phases. In each such sub-phase we randomly draw twice as many non-binding sites as there are binding sites. We then construct a Random Forest (RF) taking those non-binding sites and all binding sites of the 4 parts as input. Finally, for each instance of the validation part the majority vote of above 11 RF classifiers was taken.

#### 4. Discussion

Our results show that combining either feature  $f_{\text{JSD-t}}$  or feature  $f_{\text{JSD}}$  with the three features  $f_{\text{PSSM}}$ ,  $f_{\text{OBV}}$  and  $f_{\text{SS}}$  we have adopted from [5] clearly boosts the performance of the RF-based classifier in identifying the DNA-binding sites in proteins, where feature  $f_{\text{JSD-t}}$  generally reaches a slightly better performance than feature  $f_{\text{JSD}}$ .

Although our two new features and PSSMs are derived from MSAs, Tables 1 and 2 clearly demonstrate that these approaches carry distinct information. Thus they capture different kinds of evolutionary information. The reason for this essential difference can be explained based on the underlying algorithms. While the PSSM approach consists of statistic which indicates how likely a certain amino acids occurs at a certain position, our JSD-based approach measures the divergence of a certain distribution to a known non-binding site distribution.



The superiority of feature  $f_{\text{JSD-t}}$  to feature  $f_{\text{JSD}}$  deserves an explanation attempt. Feature  $f_{\text{JSD}}$  does not integrate any information on DNA-binding sites. Only training non-binding sites are used. In contrast, feature  $f_{\text{JSD-t}}$  additionally uses a doubly stochastic matrix gained from the training binding sites. The effect on empirical amino acid column distributions of the transformation we have devised using that matrix is the following. The empirical column probabilities of amino acids are merged, if it is very likely to co-observe them in a binding site column. Since the amino acid content of binding site columns and non-binding site columns differ, the distance between  $f_{\text{JSD-t},M}(k)$  and  $f_{\text{JSD-t},M}(k')$  is larger and more significant than the distance between  $f_{\text{JSD},M}(k)$  and  $f_{\text{JSD},M}(k')$ , where  $k$  is a binding site column of MSA  $M$ , and  $k'$  is a non-binding site column.

At first glance it is surprising that adding both feature  $f_{\text{JSD-t}}$  and feature  $f_{\text{JSD}}$  to the feature triplet ( $f_{\text{PSSM}}, f_{\text{OBV}}, f_{\text{SS}}$ ) is worse than adding feature  $f_{\text{JSD-t}}$  alone. Taking into account what we have mentioned in the preceding paragraph, it turns out that if feature  $f_{\text{JSD-t}}$  is already there, feature  $f_{\text{JSD}}$  may increase the noise.

## 5. Conclusions

In this work, we report a new sequence-based feature extraction method for the identification of DNA binding sites in proteins. For this purpose, we adopt the ideas from Capra et al. [34] and our previous studies CMF [36] and QCMF [32]. Our approach is an information theoretic method that applies the Jensen–Shannon divergence (JSD) for amino acid distributions of each site in a protein in two different ways. First, the JSD is applied to quantify the differences between observed amino acid distributions of sites and the background distribution of non-binding sites. Second, we transform the observed distributions of sites through a doubly stochastic matrix to incorporate biochemical signals of binding residues in the calculation of JSD that results in the intensification of the DNA-binding residue signals from the non-binding signals. The results of our study show that the additional usage of our new features ( $f_{\text{JSD-t}}$  or feature  $f_{\text{JSD}}$ ) in combination with existing features significantly boosts the performance of RF classifier in identifying DNA binding sites in proteins. Our results further indicate the importance of our second feature ( $f_{\text{JSD-t}}$ ) since taking into account the binding site signals in the calculation of JSD metric, the characteristics of DNA binding residues are enhanced. As a consequence, an intensification of the signal caused by DNA binding sites from non-binding sites occurs and thus the classifier achieves its improved performance.

**Acknowledgments:** We thank our colleagues Edgar Wingender, Martin Haubrock and Sebastian Zeidler for their helpful advice and insights at early stages of this project. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

**Author Contributions:** Mehmet Gültas developed the model. Stephan Waack adjusted the model together with Mehmet Gültas. Truong Khanh Linh Dang developed the model together with Mehmet Gültas, designed and implemented the tool and interpreted the results together with Cornelia Meckbach, Rebecca Tacke and Mehmet Gültas. Cornelia Meckbach and Rebecca Tacke studied the DNA binding sites in MYC-MAX protein complex. Mehmet Gültas conceived of and managed the project and wrote the final version of the manuscript. All authors have read and approved the final manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

The detailed performance of the RF classifier on different features using different cut-offs for RBscore and PreDNA datasets.

Appendix A.1. Performance Measures with Standard Error

**Table A1.** Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å. The prediction system was evaluated by five-fold cross validation.

Feature	Sensitivity ± SE(%)	Specificity ± SE(%)	MCC ± SE(%)
f <sub>PSSM</sub>	29.2 ± 2.20	96.3 ± 0.46	30.7 ± 0.95
f <sub>PSSM</sub> + f <sub>JSD</sub>	38.5 ± 3.04	94.9 ± 0.57	34.9 ± 1.7
f <sub>PSSM</sub> + f <sub>JSD-t</sub>	41.0 ± 3.23	93.9 ± 0.57	35.0 ± 1.85
f <sub>PSSM</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	41.4 ± 3.42	94.0 ± 0.51	34.8 ± 2.07
f <sub>PSSM</sub> + f <sub>SS</sub>	33.9 ± 2.32	95.8 ± 0.37	33.4 ± 1.36
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	41.6 ± 3.05	95.0 ± 0.46	37.8 ± 2.19
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	44.1 ± 3.12	94.0 ± 0.43	37.2 ± 2.37
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	43.9 ± 3.14	94.0 ± 0.40	37.0 ± 2.25
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub>	36.7 ± 2.07	96.8 ± 0.27	39.8 ± 1.58
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	42.2 ± 2.70	95.8 ± 0.42	40.9 ± 1.95
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	44.7 ± 3.05	95.0 ± 0.38	40.3 ± 1.98
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	44.4 ± 3.12	94.7 ± 0.39	39.3 ± 2.02

**Table A2.** Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å. The prediction system was evaluated by five-folds cross validation.

Feature	Sensitivity ± SE(%)	Specificity ± SE(%)	MCC ± SE(%)
f <sub>PSSM</sub>	28.6 ± 2.56	96.6 ± 0.47	35.0 ± 1.43 5
f <sub>PSSM</sub> + f <sub>JSD</sub>	39.5 ± 2.89	95.0 ± 0.55	40.7 ± 1.99
f <sub>PSSM</sub> + f <sub>JSD-t</sub>	41.8 ± 3.02	94.3 ± 0.62	41.1 ± 2.05
f <sub>PSSM</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	42.6 ± 3.25	94.2 ± 0.54	41.4 ± 2.37
f <sub>PSSM</sub> + f <sub>SS</sub>	33.4 ± 2.34	96.3 ± 0.38	38.6 ± 1.90
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	42.4 ± 2.97	95.1 ± 0.61	43.6 ± 2.43
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	44.8 ± 2.99	94.4 ± 0.56	43.8 ± 2.45
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	44.5 ± 3.04	94.4 ± 0.50	43.4 ± 2.35
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub>	33.7 ± 2.48	97.5 ± 0.35	43.1 ± 2.05
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	41.9 ± 2.89	95.8 ± 0.55	45.0 ± 2.39
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	43.9 ± 2.89	95.2 ± 0.48	45.3 ± 2.32
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	44.2 ± 2.91	94.9 ± 0.54	44.5 ± 2.24

Appendix A.2. RBscore Dataset Analysis

**Table A3.** The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f <sub>PSSM</sub>	0.458	0.974	0.476	0.866	0.460
f <sub>PSSM</sub> + f <sub>JSD</sub>	0.56	0.965	0.514	0.894	0.518
f <sub>PSSM</sub> + f <sub>JSD-t</sub>	0.597	0.957	0.511	0.899	0.523
f <sub>PSSM</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.591	0.958	0.511	0.90	0.526
f <sub>PSSM</sub> + f <sub>SS</sub>	0.512	0.97	0.501	0.878	0.476
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.581	0.96	0.511	0.899	0.520
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.611	0.953	0.508	0.903	0.526
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.613	0.953	0.509	0.902	0.528
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub>	0.517	0.976	0.534	0.896	0.528
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.58	0.967	0.54	0.907	0.543
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.612	0.963	0.546	0.910	0.551
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.601	0.962	0.531	0.909	0.546

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

**Table A4.** The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f <sub>PSSM</sub>	0.445	0.977	0.528	0.873	0.589
f <sub>PSSM</sub> + f <sub>JSD</sub>	0.553	0.968	0.579	0.899	0.643
f <sub>PSSM</sub> + f <sub>JSD-t</sub>	0.57	0.962	0.572	0.900	0.642
f <sub>PSSM</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.569	0.963	0.574	0.895	0.642
f <sub>PSSM</sub> + f <sub>SS</sub>	0.49	0.973	0.547	0.880	0.602
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.578	0.963	0.583	0.902	0.648
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.605	0.958	0.587	0.904	0.652
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.603	0.959	0.587	0.902	0.653
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub>	0.499	0.98	0.584	0.895	0.641
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.57	0.968	0.595	0.908	0.661
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.592	0.965	0.60	0.908	0.665
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.594	0.964	0.597	0.907	0.663

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

Appendix A.3. PreDNA Dataset Analysis

**Table A5.** The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f <sub>PSSM</sub>	0.378	0.977	0.41	0.840	0.391
f <sub>PSSM</sub> + f <sub>JSD</sub>	0.498	0.963	0.448	0.865	0.453
f <sub>PSSM</sub> + f <sub>JSD-t</sub>	0.543	0.953	0.445	0.869	0.451
f <sub>PSSM</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.538	0.956	0.453	0.869	0.455
f <sub>PSSM</sub> + f <sub>SS</sub>	0.393	0.975	0.417	0.847	0.402
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.501	0.966	0.461	0.872	0.463
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.545	0.959	0.465	0.876	0.468
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.523	0.958	0.449	0.875	0.465
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub>	0.428	0.977	0.458	0.867	0.451
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.511	0.97	0.488	0.885	0.488
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.539	0.962	0.475	0.888	0.488
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.539	0.961	0.47	0.886	0.488

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

**Table A6.** The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f <sub>PSSM</sub>	0.373	0.979	0.463	0.833	0.496
f <sub>PSSM</sub> + f <sub>JSD</sub>	0.485	0.962	0.495	0.858	0.540
f <sub>PSSM</sub> + f <sub>JSD-t</sub>	0.496	0.953	0.475	0.858	0.534
f <sub>PSSM</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.495	0.955	0.479	0.857	0.535
f <sub>PSSM</sub> + f <sub>SS</sub>	0.389	0.977	0.47	0.839	0.501
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.49	0.963	0.501	0.863	0.550
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.503	0.957	0.492	0.865	0.547
f <sub>PSSM</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.504	0.958	0.497	0.865	0.550
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub>	0.395	0.98	0.488	0.858	0.530
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub>	0.48	0.968	0.511	0.874	0.563
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD-t</sub>	0.506	0.962	0.51	0.873	0.560
f <sub>PSSM</sub> + f <sub>OBV</sub> + f <sub>SS</sub> + f <sub>JSD</sub> + f <sub>JSD-t</sub>	0.499	0.96	0.498	0.871	0.555

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

## References

1. Liu, B.; Wang, S.; Wang, X. DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. *Sci. Rep.* **2015**, *5*, 15479.
2. Miao, Z.; Westhof, E. Prediction of nucleic acid binding probability in proteins: A neighboring residue network based score. *Nucleic Acids Res.* **2015**, *43*, 5340–5351.
3. Si, J.; Zhang, Z.; Lin, B.; Schroeder, M.; Huang, B. MetaDBSite: A meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.* **2011**, *5* (Suppl. S1), S7.
4. Ma, X.; Guo, J.; Liu, H.D.; Xie, J.M.; Sun, X. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1766–1775.
5. Wu, J.; Liu, H.; Duan, X.; Ding, Y.; Wu, H.; Bai, Y.; Sun, X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* **2009**, *25*, 30–35.
6. Liu, B.; Xu, J.; Fan, S.; Xu, R.; Zhou, J.; Wang, X. PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. *Mol. Inform.* **2015**, *34*, 8–17.
7. Xu, R.; Zhou, J.; Wang, H.; He, Y.; Wang, X.; Liu, B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* **2015**, *9* (Suppl. S1), S10.
8. Dong, Q.; Wang, S.; Wang, K.; Liu, X.; Liu, B. Identification of DNA-binding proteins by auto-cross covariance transformation. In Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 9–12 November 2015; pp. 470–475.
9. Wei, L.; Tang, J.; Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **2016**, in press.
10. Waris, M.; Ahmad, K.; Kabir, M.; Hayat, M. Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix. *Neurocomputing* **2016**, *199*, 154–162.
11. Zhou, J.; Xu, R.; He, Y.; Lu, Q.; Wang, H.; Kong, B. PDNAsite: Identification of DNA-binding Site from Protein Sequence by Incorporating Spatial and Sequence Context. *Sci. Rep.* **2016**, *6*, 27653.
12. Jones, S.; Shanahan, H.P.; Berman, H.M.; Thornton, J.M. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* **2003**, *31*, 7189–7198.
13. Ahmad, S.; Gromiha, M.M.; Sarai, A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* **2004**, *20*, 477–486.
14. Bhardwaj, N.; Langlois, R.E.; Zhao, G.; Lu, H. Structure based prediction of binding residues on DNA-binding proteins. In Proceedings of the IEEE 27th Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS 2005), Shanghai, China, 1–4 September 2005; pp. 2611–2614.
15. Ahmad, S.; Sarai, A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinform.* **2005**, *6*, 33.
16. Kuznetsov, I.B.; Gou, Z.; Li, R.; Hwang, S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins* **2006**, *64*, 19–27.
17. Wang, L.; Brown, S.J. Prediction of DNA-binding residues from sequence features. *J. Bioinform. Comput. Biol.* **2006**, *4*, 1141–1158.
18. Wang, L.; Brown, S.J. BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.* **2006**, *34*, W243–W248.
19. Ofra, Y.; Mysore, V.; Rost, B. Prediction of DNA-binding residues from sequence. *Bioinformatics* **2007**, *23*, i347–i353.
20. Siggers, T.W.; Honig, B. Structure-based prediction of C2H2 zinc-finger binding specificity: Sensitivity to docking geometry. *Nucleic Acids Res.* **2007**, *35*, 1085–1097.
21. Tjong, H.; Zhou, H.X. DISPLAYAR: An accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.* **2007**, *35*, 1465–1477.
22. Nimrod, G.; Schushan, M.; Szilágyi, A.; Leslie, C.; Ben-Tal, N. iDBPs: A web server for the identification of DNA binding proteins. *Bioinformatics* **2010**, *26*, 692–693.
23. Wang, L.; Huang, C.; Yang, M.Q.; Yang, J.Y. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.* **2010**, *4* (Suppl. S1), S3.
24. Miao, Z.; Westhof, E. A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs. *PLoS Comput. Biol.* **2015**, *11*, e1004639.
25. Yan, J.; Friedrich, S.; Kurgan, L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief. Bioinform.* **2015**, *17*, 88–105.

26. Yan, C.; Terribilini, M.; Wu, F.; Jernigan, R.L.; Dobbs, D.; Honavar, V. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinform.* **2006**, *7*, 262.
27. Hwang, S.; Gou, Z.; Kuznetsov, I.B. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* **2007**, *23*, 634–636.
28. Huang, Y.F.; Huang, C.C.; Liu, Y.C.; Oyang, Y.J.; Huang, C.K. DNA-binding residues and binding mode prediction with binding-mechanism concerned models. *BMC Genom.* **2009**, *10* (Suppl. S3), S23.
29. Wong, K.C.; Li, Y.; Peng, C.; Moses, A.M.; Zhang, Z. Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res.* **2015**, *43*, 10180–10189.
30. Wang, L.; Yang, M.Q.; Yang, J.Y. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genom.* **2009**, *10* (Suppl. S1), S1.
31. Eggeling, R.; Roos, T.; Myllymäki, P.; Grosse, I. Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinform.* **2015**, *16*, doi:10.1186/s12859-015-0797-4.
32. Gültas, M.; Düzgün, G.; Herzog, S.; Jäger, S.J.; Meckbach, C.; Wingender, E.; Waack, S. Quantum coupled mutation finder: Predicting functionally or structurally important sites in proteins using quantum Jensen–Shannon divergence and CUDA programming. *BMC Bioinform.* **2014**, *15*, 96.
33. Fischer, J.; Mayer, C.E.; Söding, J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **2008**, *24*, 613–620.
34. Capra, J.A.; Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **2007**, *23*, 1875–1882.
35. Grosse, I.; Bernaola-Galván, P.; Carpena, P.; Román-Roldán, R.; Oliver, J.; Stanley, H.E. Analysis of symbolic sequences using the Jensen–Shannon divergence. *Phys. Rev. E* **2002**, *65*, 041905.
36. Gültas, M.; Haubrock, M.; Tüysüz, N.; Waack, S. Coupled mutation finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. *BMC Bioinform.* **2012**, *13*, 225.
37. Li, T.; Li, Q.Z.; Liu, S.; Fan, G.L.; Zuo, Y.C.; Peng, Y. PreDNA: Accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. *Bioinformatics* **2013**, *29*, 678–685.
38. Krall, A.; Brunn, J.; Kankanala, S.; Peters, M.H. A simple contact mapping algorithm for identifying potential peptide mimetics in protein–protein interaction partners. *Proteins* **2014**, *82*, 2253–2262.
39. Nair, S.K.; Burley, S.K. X-ray structures of Myc-Max and Mad-Max recognizing DNA: Molecular bases of regulation by proto-oncogenic transcription factors. *Cell* **2003**, *112*, 193–205.
40. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
41. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173–175.
42. Cappellini, V.; Sommer, H.J.; Bruzda, W.; Zyczkowski, K. Random bistochastic matrices. *J. Phys. A Math. Theor.* **2009**, *42*, 36.
43. Bartlett, G.J.; Porter, C.T.; Borkakoti, N.; Thornton, J.M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **2002**, *324*, 105–121.
44. Panchenko, A.R.; Kondrashov, F.; Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* **2004**, *13*, 884–892.
45. Janda, J.O.; Busch, M.; Kück, F.; Porfenenko, M.; Merkl, R. CLIPS-1D: Analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure. *BMC Bioinform.* **2012**, *13*, 55.
46. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
47. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.
48. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.

