Evaluierung des phylogenetischen Footprintings und dessen Anwendung zur verbesserten Vorhersage von Transkriptionsfaktor-Bindestellen

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von

Tilman Sauer

aus Gifhorn

Göttingen 2006

D 7
Referent: Prof. Dr. Stephan Waack
Korreferent: Prof. Dr. Edgar Wingender
Tag der mündlichen Prüfung: 11.07.2006

Danksagung

An dieser Stelle möchte ich allen danken, die mich während dieser Arbeit unterstützt haben. Mein ganz besonderer Dank gilt Professor Edgar Wingender, der diese Arbeit betreut und überhaupt erst ermöglicht hat. Er bereicherte diese Arbeit sehr durch viele Anregungen, fruchtbare Diskussionen, seine umfassende wissenschaftliche Erfahrung und seine Expertise im Bereich der Transkriptionsregulation. Professor Stephan Waack danke ich herzlichst dafür, Mentor und Erstgutachter dieser Arbeit zu sein. Er unterstützte mich sehr durch viele inspirierende Diskussionen und sein umfangreiches Wissen über Hidden-Markov-Modelle. Das interdisziplinäre Arbeiten mit ihm war eine große Bereicherung.

Ich danke der gesamten Abteilung Bioinformatik für ihre Unterstützung und ihre Diskussionsbereitschaft. Für ihre Hilfsbereitschaft bei den kleineren Problemen des Programmieralltags danke ich insbesondere Torsten Crass, Jürgen Dönitz, Torsten Schöps und Knut Schwarzer. Ein großer Dank geht an Martin Haubrock, der das Büro mit mir geteilt hat und immer ein offenes Ohr hatte, wenn es Probleme oder Ergebnisse zu diskutieren gab. Fabian und Dilei danke ich für ihre Mühen während ihrer Bachelor-Arbeit. Für die Hilfe in administrativen Angelegenheiten danke ich Carmen Modrok und Doris Waldmann.

Anders (Maat), Bernd (Disco), Jens (Lan), Niels (Retro), Martin, Sven (Solo), Than-Duc (Duke), Tim (Mokele) und Torben (Torbs) danke ich für ihre Freundschaft während dieser Zeit.

Anne danke ich von ganzem Herzen für das kritische Lesen der Arbeit und viele Diskussionen, für ihre Unterstützung und dafür, dass sie für mich da war.

Mein tiefer Dank gilt meiner Familie, insbesondere meinen Schwestern Johanna und Laura, und meinen Eltern, die mich immer bedingungslos auf meinem Weg unterstützt haben.

Inhaltsverzeichnis

Abkürzungsverzeichnis

A	Adenin
API	application programming interface
BLAST	Basic Local Alignment Search Tool
C	Cytosin
CE	composite element
CSS	core similarity score
DNA	deoxyribonucleic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
G	Guanin
Gb	Giga-Basenpaare
GFF	General Feature Format
GO	gene ontology
GRN	genregulatorisches Netzwerk
HMM	Hidden-Markov-Modell
HSP	high-scoring segment pair
IUPAC	International Union of Pure and Applied Chemistry
kb	Kilo-Basenpaare
KS	Konserviertheit-Schwellenwert
LS	linearer Score
Mb	Mega Basenpaare
MSS	matrix similarity score
NHR	nuclear hormone receptor
OSP	orthologes Sequenzpaar
PI	prozentuale Identität
PSSM	positions-spezifische Scoring-Matrix
RF	regulatory feature
ROC	receiver operating characteristic

T	Thymin
TBP	TATA-Bindeprotein
TF	Transkriptionsfaktor
TFBS	Transkriptionsfaktor-Bindestelle
TLS	Translationsstartstelle
TSS	Transkriptionsstartstelle
UCR	ultra-conserved region
UTR	untranslated region
WGA	whole genome alignment

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Seit einigen Jahren ist die vollständige Sequenz des Humangenoms bekannt, welche aus ca. 3.2 Milliarden Nukleotiden besteht. Die riesige Sequenzinformation steht allerdings in großer Diskrepanz zu dem bisher noch geringen Verständnis dieser Information. Das Humangenom besitzt zwischen 20000 und 25000 Gene, die für Proteine codieren, und RNA-Gene unterschiedlicher Funktion, aber ungefähr 95% des Humangenoms werden nicht transkribiert. In dieser Datenflut sind jedoch alle Informationen zur Strukturorganisation und Transkriptionsregulation des Genoms verborgen. Eine der größten Herausforderungen der heutigen Zeit ist daher die Gewinnung von Wissen aus den zur Verfügung stehenden riesigen Sequenzdatenmengen.

Das Verständnis der Transkriptionsregulation ist von entscheidender Bedeutung, um die komplexen Vorgänge innerhalb einer Zelle analysieren zu können. Für die Entschlüsselung der genregulatorischen Netzwerke von Zellen benötigt man zuallererst die Kenntnis über die regulatorischen Elemente der Gene im Netzwerk. Bestimmte Proteine, die sogenannten Transkriptionsfaktoren (TFs), binden an solche regulatorische Elemente, die daher auch Transkriptionsfaktor-Bindestellen (TFBSs) genannt werden, und beeinflussen die Effizienz der Transkription des regulierten Gens. TFBSs können mit verschiedenen Methoden experimentell identifziert werden. Eine häufig verwendete Technik ist das „DNA-Footprinting“, bei dem die exakte Sequenz, an die ein Protein bindet, durch chemischen oder enzymatischen Abbau der DNA bestimmt wird. Dabei bleiben nur die Sequenzabschnitte, an die ein Protein bindet, intakt und können anschließend sequenziert werden. Diese und andere experimentelle Techniken sind jedoch sehr zeit- und kostenintensiv.

Eine andere Möglichkeit zur Bestimmung von TFBSs sind bioinformatische Methoden, die alleinig auf vorhandenen Sequenzinformationen beruhen. Regulatorische Elemente sind kurz und degeneriert, daher ist die Chance, dass ein bestimmtes Sequenzmuster zufällig auftaucht, relativ hoch, was die zuverlässige bioinformatische Detektion von

TFBSs erschwert. Um das Signal-zu-Rausch-Verhältnis bei dieser Suche zu verbessern, wird häufig die Sequenz-Konserviertheit zwischen zwei oder mehreren Spezies als Mittel genutzt, um falsch positive Vorhersagen zu minimieren. Dieser Ansatz wird in Anlehnung an die Technik des DNA-Footprinting auch „phylogenetisches Footprinting“ genannt. Phylogenetisches Footprinting bedient sich des Vergleichs nicht-codierender Regionen, die stromaufwärts orthologer Gene liegen, um regulatorische Bereiche zu identifizieren. Phylogenetisches Footprinting basiert auf der Annahme, dass funktionelle Bereiche in nicht-codierenden Sequenzen einem höheren evolutionären Druck unterliegen als nicht-funktionelle Bereiche. Mutationen in regulatorischen Elementen, die deren Funktion beeinträchtigen oder außer Kraft setzen, werden daher durch Selektion aus einer Population entfernt. Als Folge dessen sollten regulatorische Bereiche im Laufe der Evolution weniger Mutationen anhäufen als nicht-funktionelle Bereiche. Phylogenetisches Footprinting ist eine häufig genutzte Methode, wobei zur Detektion regulatorischer Elemente im Humangenom Sequenz-Vergleiche mit den Genomen der Maus oder der Ratte gängige Praxis sind.

Im Rahmen dieser Arbeit wurde der Ansatz des phylogenetischen Footprintings evaluiert. Dazu wurde untersucht, inwiefern experimentell verifizierte TFBSs durch Sequenzvergleiche zwischen Mensch und Maus oder Ratte detektiert werden können, um den Ansatz zu kalibrieren und einzuschätzen. Drei Hauptgesichtspunkte wurden analysiert: die Sicherstellung der orthologen Beziehung der zu vergleichenden Sequenzen, der Einfluss des benutzten Alignment-Programms auf die Ergebnisse und eine geeignete Definition der Konserviertheit, sodass einerseits möglichst viele TFBSs erkannt werden, aber andererseits der Anteil an falsch positiven Vorhersagen gering gehalten wird.Da TFBSs des gleichen TF zum Teil erheblich in ihrer Sequenz variieren können, wurde die Konserviertheit zusätzlich von der Sequenzebene abstrahiert und auf einer Musterebene untersucht. Um den Einfluss des evolutionären Abstands der zu vergleichenden Spezies auf den Erfolg des phylogenetischen Footprintings einzuschätzen, wurde zudem verglichen, inwiefern Mensch-TFBSs durch Sequenz-Vergleiche mit den Genomen des Hunds oder der Kuh detektiert werden können.

Die Ergebnisse der Evaluation des phylogenetischen Footprintings wurden genutzt, um regulatorische Elemente besser vorherzusagen, als dies mit bisher etablierten Verfahren möglich ist. Ein gängiges Verfahren basiert auf der Suche nach bestimmten Sequenzmustern mit sogenannten positions-spezifischen Scoring-Matrizen (PSSMs). Da die Informationen aus dem phylogenetischen Footprinting einen davon unabhängigen Hinweis auf die Existenz einer TFBS liefern, sollte die Kombination aus einer PSSM-basierten Vorhersage von TFBSs und phylogenetischem Footprinting die Anzahl falsch positiver Vorhersagen

verringern. Zu diesem Zweck wurde ein Hidden-Markov-Modell (HMM) zur Vorhersage von TFBSs entworfen, das die Musterinformation einer PSSM und die Konserviertheit zwischen zwei Spezies auf synergistische Weise verknüpft. Die Vorhersagen des HMM wurden mit PSSM-basierten Vorhersagen verglichen und anhand bekannter TFBSs überprüft.

Diese Arbeit gibt in Kapitel 2 zunächst einen Überblick über die biologischen Grundlagen der Genregulation und bioinformatische Methoden zur Detektion von TFBSs. In Kapitel 3 werden die benutzten Datenbanken und Programme, die Methoden zur Analyse der Daten und das entwickelte HMM beschrieben. Die Ergebnisse aller Untersuchungen werden in Kapitel 4 vorgestellt und diskutiert.

Kapitel 2

Hintergrund und Übersicht

2.1 Biologische Grundlagen der Genregulation

Im Folgenden wird kurz auf die biologischen Grundlagen der Genregulation eingegangen.

2.1.1 DNA

Die Desoxyribonukleinsäure, meist nach der englischen Bezeichnung „deoxyribonucleic acid“ mit DNA abgekürzt, ist ein langes Polymer von Nukleotiden. Sie enthält genetische Informationen, die die biologische Entwicklung aller zellulären Formen des Lebens spezifizieren. Die DNA ist eine Doppelhelix aus antiparallelen Strängen (siehe Abbildung 2.1), deren Nukleotide durch 5’-3’-Phosphodiesterbindungen verknüpft sind.

Ein Nukleotid besteht aus dem Zucker Desoxyribose, einer Phosphatgruppe und einer der vier heterozyklischen aromatischen Nukleobasen: Adenin (A) , Cytosin (C) , Guanin (G) und Thymin (T). Es gibt zwei Arten von Nukleobasen, die Purin- und die Pyrimidinbasen. Pyrimidine (C und T) besitzen einen sechsgliedrigen Ring, Purine (A und G) einen fünf- und einen sechsgliedrigen Ring, die miteinander verbunden sind.

Das polare Zucker-Phosphat-Rückgrat bildet die Außenseite der Doppelhelix, die unpolaren Purin- und Pyrimidinbasen sind im Inneren paarweise gestapelt. Die Basenpaare tragen in zweierlei Hinsicht zur thermodynamischen Stabilität der Doppelhelix bei. Zum einen bilden sich zwischen komplementären Basen (A ist komplementär zu T und G zu C) Wasserstoffbrückenbindungen aus, wobei zwischen A und T zwei und zwischen G und C drei Wasserstoffbrücken ausgebildet werden (siehe Abbildung 2.2). Zum anderen gibt es eine Wechselwirkung zwischen den π-Elektronen der Basenpaare, die in der sogenannten „hydrophoben Stapelung“ („π-Stacking“) der Basen resultiert.

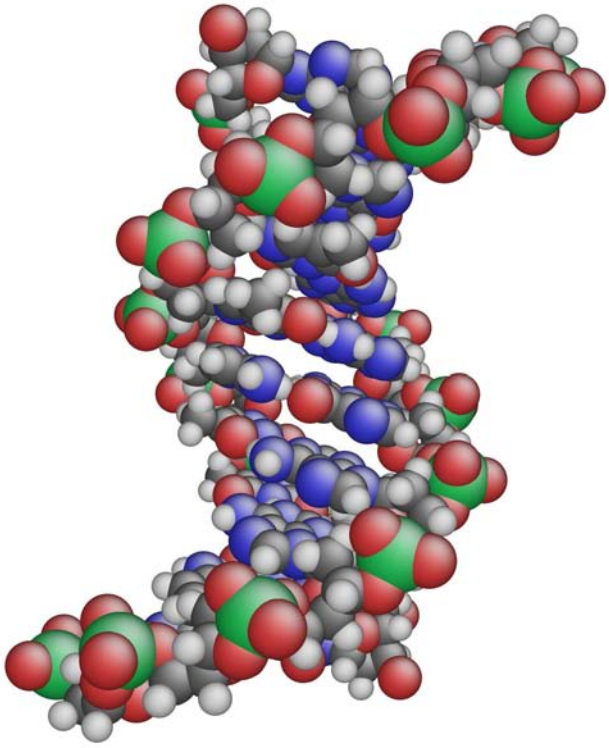

Abbildung 2.1: Molekulare Struktur der DNA. Die DNA ist eine Doppelhelix aus zwei antiparallelen Strängen, deren Rückgrat aus abwechselnd angeordneten Desoxyribose- und Phosphatmolekülen besteht, die über 5'-3'-Phosphodiesterbindungen verknüpft sind. Im Inneren der Helix befi nden sich die Nukleobasen. Die Abbildung wurde mit AtomEye (Li, 2003) erzeugt.

2.1.2 Gene und Promotoren

Gene sind die grundlegenden biologischen Einheiten der Vererbung. Die meisten Gene sind DNA-Abschnitte, die für ein Protein codieren. Ein eukaryotisches Gen besteht aus der zu transkribierenden Sequenz, die sich aus den Exons und den dazwischenliegenden Introns zusammensetzt, und aus den Bereichen, die der zu transkribierenden Sequenz vorausgehen und folgen (siehe Abbildung 2.3).

Die Transkription eines Gens liefert die „prä-messenger-RNA". Durch das sogenannte „Splicing" werden die Introns aus der transkribierten Sequenz entfernt und die Exons zur „messenger-RNA" (mRNA) verbunden, die für ein Protein codiert. Die Nukleotidsequenz der mRNA codiert die Abfolge der Aminosäuren („genetischer Code"), aus denen während der Translation das entsprechende Polypeptid gebildet wird.

Der Promotor eines Gens ist die DNA-Sequenz, die die Transkription des Gens reguliert und typischerweise stromaufwärts des Gens liegt. Die meisten Promotoren enthalten mehrere *cis*-regulatorische DNA-Elemente, an die *trans*-agierende, regulatorische Protei-

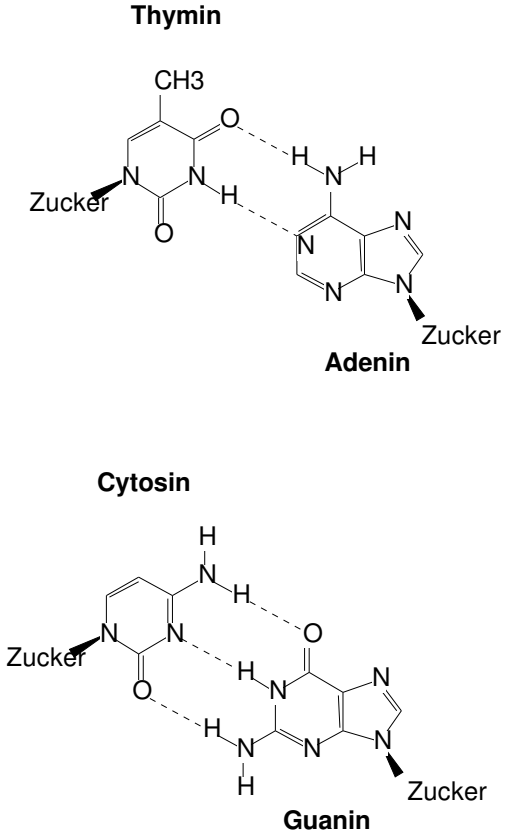

Abbildung 2.2: Komplementäre Basenpaarungen der DNA. Zwischen A und T werden zwei, zwischen C und G drei Wasserstoffbrücken ausgebildet.

ne, die Transkriptionsfaktoren (TFs) genannt werden, binden. Die *cis*-regulatorischen Elemente werden daher auch Transkriptionsfaktor-Bindestellen (TFBSs) genannt (siehe 2.1.3, S. 13).

Der Promotor eines Gens lässt sich grob in drei Bereiche aufteilen: den Kernpromotor, den proximalen und den distalen Promotor. Der Kernpromotor ist der Bereich von ca. Position -35 bis zur Transkriptionsstartstelle (TSS) . Der proximale Promotor reicht bis etwa zur Position -250 stromaufwärts der TSS und der distale Promotor umfasst alle TFBSs, die noch weiter stromaufwärts liegen (Lewin, 2002).

Ein Promotor ist modular aufgebaut, d.h. er enthält mehrere kurze *cis*-regulatorische Sequenzelemente, die von TFs (siehe 2.1.3, S. 13) erkannt werden. Die Sequenzabschnitte zwischen den TFBSs sind an sich irrelevant, können aber die Funktion von Abstandhaltern („spacer“) übernehmen. Verschiedene Promotoren enthalten unterschiedliche Kombinationen dieser *cis*-regulatorischen Elemente. Dies ermöglicht eine gezielte Regulation einzelner Gene, die unter z.T. sehr unterschiedlichen Bedingungen oder z.B. auch nur in bestimmten Zelltypen benötigt werden.

Insgesamt sind eukaryotische Promotoren extrem divers. Sie variieren sehr stark in ihrer

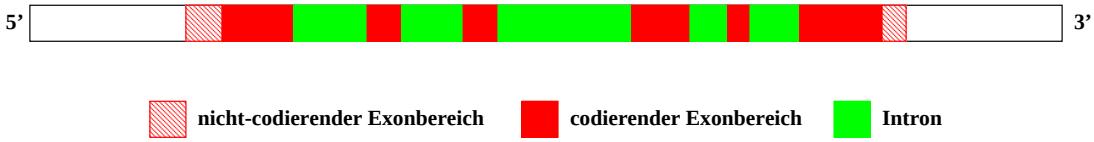

Abbildung 2.3: Struktur eines Gens. Ein Gen setzt sich aus den nicht-codierenden und codierenden Exons, den dazwischenliegenden Introns und den Bereichen stromaufwärts und stromabwärts zusammen. In diesem Beispiel besteht das Gen aus fünf Introns und sechs Exons, von denen das erste und letzte Exon nicht-codierende Bereiche enthalten.

Sequenz und sind daher schwierig zu charakterisieren. Beispielsweise ist es problematisch die Länge eines Promotors festzulegen, denn regulatorische Elemente können selbst im Abstand von mehreren Kilobasen zur TSS vorkommen. Der Unterschied zwischen Promotoren und sogenannten „Enhancern“ ist fließend. Enhancer können in variablen Abständen sowohl stromauf- und stromabwärts des Promotors liegen. Enhancer sind wie Promotoren modular aus *cis*-regulatorischen Elementen aufgebaut. Sie wirken in der Regel positions- und orientierungsunabhängig und regulieren wie Promotoren die Effizienz der Transkription von Genen.

Der Großteil der regulatorischen Elemente liegt allerdings im Bereich der ersten 300 bp direkt vor der TSS. Zusammenfassend lässt sich sagen, dass der Promotor dem DNA-Bereich entspricht, der alle TFBSs eines Gens enthält und so die Transkription des Gens mit der notwendigen Effizienz und unter geeigneter Kontrolle ermöglicht.

2.1.3 Transkriptionsfaktoren und Transkriptionsfaktor-Bindestellen

Jedes Protein, das für die Initiation der Transkription benötigt wird, selbst aber kein Bestandteil der RNA-Polymerase ist, wird als TF definiert. Die meisten TFs interagieren über eine DNA-bindende Domäne direkt mit der DNA (siehe Abbildung 2.4).

Die Transkription in Eukaryoten wird von zwei Klassen von TFs initiiert, den sogenannten allgemeinen und den spezifischen TFs. Die allgemeinen TFs binden an den Kernpromotor, der nahe der TSS liegt, und bilden zusammen mit der RNA-Polymerase den Initiationskomplex. Sie sind für die Ablaufmechanik der Enzymbindung der RNA-Polymerase an die DNA und die Transkriptionsinitiation notwendig. Ein regulatorischer Einfluss geht von den allgemeinen TFs nicht aus.

Zwei der am besten charakterisierten Kernpromotor-Elemente sind die TATA-Box und das Pyrimidin-reiche Inr-Element. Die TATA-Box liegt ca. 25 bp stromaufwärts der TSS. Das TATA-Bindeprotein (TBP) erkennt die A/T-reiche Sequenz der TATA-Box durch eine Interaktion mit der kleinen Furche der DNA und bewirkt eine ausgeprägte Strukturverän-

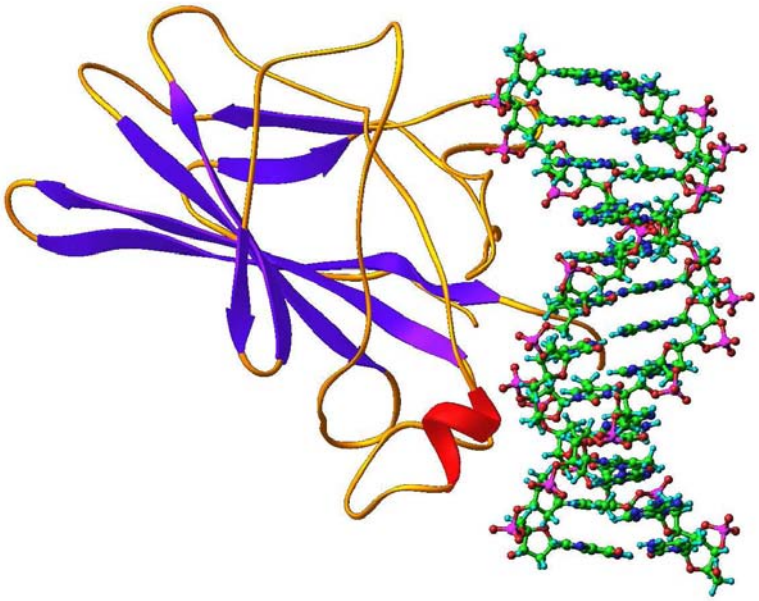

Abbildung 2.4: Bindung eines Proteins der NFAT-Familie („nuclear factor of the activated T cell“) an die DNA (Quelle: http://www-nmr.cabm.rutgers.edu/photogallery/structures/gif/a66.gif).

derung der Doppelhelix. Das Inr-Element befindet sich direkt an der TSS und hat die Konsensussequenz 5’-YYNWYY-3’ (dabei steht Y für C oder T, N für ein beliebiges Nukleotid und W für A oder T). Viele eukaryotische Promotoren weisen eine TATA-Box und/oder das Inr-Element auf, es gibt jedoch auch einige Promotoren, die keines dieser beiden Elemente besitzen.

Die spezifischen TFs binden an regulatorische DNA-Elemente im proximalen und distalen Promotor, die in unterschiedlichen Abständen zur TSS vorkommen. Sie besitzen eine sequenzspezifische DNA-bindende Domäne sowie eine *trans*-aktivierende Region, die mit anderen TFs interagiert. Die spezifischen TFs können anhand ihrer Funktionsweise in zwei Untergruppen unterteilt werden. Die erste Untergruppe sind die konstitutiven Faktoren, die bestimmte Konsensussequenzen stromaufwärts („upstream“) der TSS erkennen. Diese TFs sind ubiquitär, d.h. die Aktivität dieser TFs wird nicht reguliert, und binden an jeden Promotor, dessen DNA-Sequenz eine passende TFBS enthält. Sie regulieren die Effizienz des Transkriptionsstarts. Jeder Promotor benötigt einen für ihn charakteristischen Satz solcher TFs, damit die volle Expressionsstärke des regulierten Gens erreicht werden kann. Die zweite Untergruppe sind die induzierbaren Faktoren, die grundsätzlich wie die konstitutiven Faktoren agieren, jedoch im Gegensatz zu diesen eine regulatorische Funktion ausüben. Sie werden nur zu bestimmten Zeiten oder in bestimmten Gewebetypen synthe-

tisiert oder aktiviert, und kontrollieren dadurch sich räumlich und zeitlich ändernde Transkriptionsmuster. Die Sequenzabschnitte, an die sie binden, nennt man Reaktions- oder Response-Elemente. Die Unterscheidung zwischen konstitutiven und induzierbaren Faktoren sollte allerdings nicht als strikte Trennung angesehen werden, da es TFs gibt, die keiner Untergruppe eindeutig zugeordnet werden können. Die konstitutiven und die induzierbaren TFs üben ihre Wirkung aus, indem sie entweder direkt oder durch Mediatoren mit der RNA-Polymerase im Initiationskomplex interagieren.

Die spezifischen TFs weisen eine beachtliche Flexibilität in ihrem Bindungsverhalten auf, d.h. ein und derselbe TF kann an unterschiedliche DNA-Sequenzen binden (Müller, 2001). Diese lokalen Unterschiede in der DNA-Sequenz können durch Veränderungen der Konformation von Seitenketten des Proteins (Chen et al., 2000) oder durch Veränderungen der Struktur der Hydrathülle kompensiert werden (Schwabe, 1997), um eine Bindung des TF an die DNA zu ermöglichen. Eine Reihe von TFs sind Dimere, wobei jedes Monomer nur mit einer Hälfte der TFBS interagiert. Der Abstand zwischen den beiden Hälften der TFBS kann dabei variieren. Übersichten über verschiedene Strukturgruppen von TFs geben z.B. Wingender (1997) und Luscombe et al. (2000). Durch die Kombination von mehreren DNA-bindenden Proteinen in einem Komplex wird die Sequenzspezifität jeder individuellen Bindung gelockert, da die gesamte Kontaktfläche der Bindung größer ist. Dadurch ist es für den Komplex möglich, ein breiteres Spektrum an TFBSs zu erkennen (Chen, 1999).

Die beschriebene hohe Variabilität der TFBSs ein und desselben TF ist biologisch sinnvoll, da Gene mit unterschiedlichen Stärken exprimiert werden müssen. Dies kann dadurch erreicht werden, dass TFBSs unterschiedliche Sequenzen und damit auch unterschiedlich hohe Affinitäten zu einem bestimmten TF haben. Allerdings erschwert dies die zuverlässige bioinformatische Vorhersage von TFBSs.

2.1.4 Genregulatorische Netzwerke

Ein einfaches genregulatorisches Netzwerk (GRN) besteht aus mehreren Komponenten: einem oder mehreren Signalwegen, TFs, die die eingehenden Signale aufnehmen, einigen Zielgenen und deren TFBSs sowie der mRNA und den Proteinen, die aus der mRNA produziert werden.

Die Signalwege empfangen intra- und/oder extrazelluläre Signale und leiten sie an TFs weiter, die nun wiederum an TFBSs binden und die Expression der dazugehörigen Gene regulieren. Dabei interagieren TFs oft miteinander und bilden Multiprotein-Komplexe, die an die DNA gebunden sind. Diese Komplexbildung ermöglicht eine sehr hohe Spezifität der Genregulation. Jedes Gen ist durch seine *cis*-regulatorischen Sequenzelemente cha-

rakterisiert. Gene, die im Zusammenspiel eine bestimmte Funktion in der Zelle erfüllen, sind oft koreguliert und teilen ähnliche *cis*-regulatorische DNA-Sequenzen. Wenn TFs mit diesen TFBSs assoziieren, können sie die Transkription des regulierten Gens induzieren oder reprimieren. Dies ermöglicht der Zelle, auf intra- und/oder extrazelluläre Signale zu reagieren und ihr Expressionsprofil anzupassen.

Zusammenfassend lässt sich sagen, dass GRNs das Expressionsniveau jedes Gens im Genom verändern, indem sie kontrollieren, ob und in welchem Umfang ein Gen in mRNA transkribiert wird. Die produzierten mRNA-Transkripte dienen dann der Protein-Synthese während der Translation.

Eine klassische bioinformatische Fragestellung ist die Erstellung eines GRN basierend auf den Vorhersagen von TFBSs in Promotoren von potentiell koregulierten Genen. Koregulierte Gene können beispielsweise an einem Reaktionspfad beteiligt sein, der aus mehreren Schritten besteht, wobei jedes Genprodukt an einem dieser Schritte beteiligt ist. Auch eine Koexpression in einem Microarray-Experiment kann ein Hinweis auf eine Koregulation der entsprechenden Gene sein. Ein medizinisches Anwendungsziel ist die Identifikation eines Schlüsselproteins oder -moleküls in einem GRN, das z.B. während einer Infektion aktiviert wird, um gezielte therapeutische Maßnahmen ergreifen zu können. Die möglichst zuverlässige Vorhersage von TFBSs ist essentiell für die Erstellung eines GRN, da die TFBSs die Basis für die Netzwerkkonstruktion darstellen.

2.2 Bioinformatische Methoden zur Identifizierung von Transkriptionsfaktor-Bindestellen

Experimentelle Methoden zur exakten Lokalisierung von TFBSs wie z.B. „direct gel shift“ und „DNAse I footprinting“ (Rippe et al., 2001) sind sehr zeit- und kostenintensiv. Neuere Methoden wie „ChIP-chip“ (Buck und Lieb, 2005) erlauben es zwar, TFBSs für ganze Chromosomen zu bestimmen, allerdings kann dabei die Lokalisation von TFBSs nur auf Fragmente einer Länge von ca. 1kb eingegrenzt werden.

Eine sehr große Aufgabe in der Bioinformatik ist daher die Entwicklung und Anwendung von Methoden zur rechnergestützten Vorhersage von TFBSs. Dabei gilt es, möglichst viele funktionelle TFBSs zu erkennen, und gleichzeitig möglichst wenig falsche Vorhersagen zu machen. In den nächsten Abschnitten werden Methoden zur Vorhersage von TFBSs vorgestellt.

2.2.1 „String matching“ und „consensus matching“

Um potentielle neue TFBSs in DNA-Sequenzen zu suchen, bedient man sich sehr häufig bekannter TFBSs. Der einfachste Ansatz ist das sogenannte „string matching“: Hier wird die Sequenz einer bekannten TFBS benutzt, um diese in einer anderen Sequenz wiederzufinden. Da TFs jedoch eine z.T. erhebliche Variabilität ihrer DNA-Bindungs-Spezifität aufweisen, können viele TFBS mit dieser Methode nicht erkannt werden, d.h. man erhält viele falsch negative Ergebnisse.

Aufgrund dieser Variabilität werden die Präferenzen für DNA-Bindestellen von TFs gewöhnlich mit einer Konsensussequenz (Waterman et al., 1984) beschrieben. Konsensussequenzen spiegeln die Variabilität innerhalb einer Gruppe von bekannten TFBSs eines TF wider, indem sie (partiell) entartete Nukleotide erlauben. Beim sogenannten „consensus matching“ wird eine Sequenz nach einer Übereinstimmung mit der Konsensussequenz durchsucht, wodurch weniger falsch negative Ergebnisse als beim „string matching“ erhalten werden. Allerdings steigt die Zahl falsch positiver Ergebnisse. Wie man eine Konsensussequenz definiert, ist jedoch etwas willkürlich und nicht immer eindeutig. Die Wahl der Methode zur Bestimmung der Konsensussequenz (Cavener, 1987; Choo et al., 1991; Day und McMorris, 1992a) beeinflusst die Anzahl der Mismatches in der Konsensussequenz sowie deren erlaubte Mehrdeutigkeit und damit auch die Zahl falsch negativer und falsch positiver Ergebnisse bei der Vorhersage neuer TFBSs. Tabelle 2.1 zeigt z.B. den sogenannten „IUPAC-Code“ der „International Union of Pure and Applied Chemistry“ (IUPAC) für entartete Nukleotide zur Erzeugung von Konsensussequenzen.

Tabelle 2.1: IUPAC-Code für entartete Nukleotide

Nukleotide	IUPAC-Code
A oder G	R
C oder T	Y
A oder C	M
G oder T	K
C oder G	S
A oder T	W
A, C oder T	H
C, G oder T	B
A, C oder G	V
A, G oder T	D
A, C, G oder T	N

Eine IUPAC-Konsensussequenz wird nach folgenden Regeln erzeugt: Zuerst werden die Sequenzen der bekannten TFBSs eines TF aligniert und für jede Position die Häufigkeiten der einzelnen Nukleotide gezählt. Ist die relative Häufigkeit eines Nukleotids an einer Position größer als 50% und mindestens zweimal so hoch wie die des zweithäufigsten Nukleotids an dieser Position, wird dieses Nukleotid als Konsensus-Nukleotid bezeichnet. Haben an einer Position zwei Nukleotide zusammen eine relative Häufigkeit von mindestens 75% und trifft die erste Regel nicht zu, so wird der IUPAC-Code für die entsprechenden zwei Nukleotide als Konsensus-Nukleotid gewählt. Wenn an einer Position drei Nukleotide zusammen eine relative Häufigkeit von 100% haben und die ersten beiden Regeln nicht zutreffen, wird der IUPAC-Code für die entsprechenden drei Nukleotide als Konsensus-Nukleotid angenommen. Trifft an einer Position keine der genannten Regeln zu, wird der IUPAC-Code 'N' als Konsensus-Nukleotid gewählt. Ein Beispiel für die Konstruktion einer IUPAC-Konsensussequenz nach diesen Regeln findet sich in Abbildung 2.5.

Es ist relativ einfach, eine Konsensussequenz zu finden, die eine Gruppe von TFBSs repräsentiert. Das eigentliche Problem ist aber, eine Konsensussequenz zu bestimmen, mittels der man optimal neue TFBSs vorhersagen kann. Day und McMorris (1992b) haben mehrere Methoden zur Konstruktion von Konsensussequenzen verglichen und deren Vor- und Nachteile diskutiert. Sie kamen zu dem Schluss, dass sich funktionelle TFBSs i.A. aufgrund vieler falsch positiver Vorhersagen nur schwerlich mittels Konsensussequenzen vom genomischen Hintergrund trennen.

```
                    TFBS 1:   GTGACTCAG
                    TFBS 2:   ATGACTCAG
                    TFBS 3:   ATGACATCA
                    TFBS 4:   CTGACTCAT
                    TFBS 5:   ATGACTAAC
                    TFBS 6:   GTGACGAAA

IUPAC-Konsensussequenz:   RTGACTMAN
```

Abbildung 2.5: Beispiel der Erzeugung einer IUPAC-Konsensussequenz. Aus den aneinander ausgerichteten TFBS-Sequenzen wird nach bestimmten Regeln die IUPAC-Konsensussequenz erzeugt (siehe auch Tabelle 2.1). Diese enthält zwei doppelt und eine vierfach entartete Position, d.h. es existieren 16 verschiedene mögliche Sequenzen, die mit dieser IUPAC-Konsensussequenz übereinstimmen.

2.2.2 Positions-spezifische Scoring-Matrizen

Eine Verallgemeinerung des Konzeptes der Konsensussequenz stellt eine sogenannte „positions-spezifische Scoring-Matrix“ (PSSM) dar, mit der man die Variabilität von TFBSs besser erfassen kann. Dazu werden experimentell bekannte TFBSs eines TF aligniert und in einer Matrix, der sogenannten PSSM, auf folgende Art und Weise zusammengefasst: Diese Matrix hat die Länge der TFBSs und jede Spalte repräsentiert die Häufigkeiten der entsprechenden Nukleotide an dieser Position. Abbildung 2.6 zeigt die Konstruktion einer PSSM basierend auf den TFBSs aus Abbildung 2.5.

```
Position:      1 2 3 4 5 6 7 8 9

TFBSs:         G T G A C T C A G
               A T G A C T C A G
               A T G A C A T C A
               C T G A C T C A T
               A T G A C T A A C
               G T G A C G A A A

PSSM:      A   3 0 0 6 0 1 2 5 2
           C   1 0 0 0 6 0 3 1 1
           G   2 0 6 0 0 1 0 0 2
           T   0 6 0 0 0 4 1 0 1
```

Abbildung 2.6: PSSM-Konstruktion mittels bekannter alignierter TFBSs. Jeder Eintrag in der Matrix repräsentiert die Häufigkeit des entsprechenden Nukleotids an der jeweiligen Position im Alignment der TFBSs.

Stormo et al. (1982a) untersuchten Stellen für den Start der Translation („translation initiation sites“) in *E. coli* und führten erstmalig PSSMs (Stormo et al., 1982b) ein, um diese

vorherzusagen. Die erhaltenen Vorhersagen waren genauer als alle, die mit Konsensussequenzen erhalten wurden. PSSMs nutzen einen größeren Anteil der Sequenzinformation als Konsensussequenzen und sind daher akkurater in der Vorhersage neuer potentieller TFBSs. Während Konsensussequenzen zwar Mismatches erlauben, können PSSMs diese zusätzlich in Bezug auf die Nukleotidhäufigkeitverteilung an der entsprechenden Position bewerten. Da alle Positionen in der PSSM unabhängig voneinander betrachtet werden, gehen jedoch, wie auch bei der Konsensussequenz, Informationen über Korrelationen zwischen bestimmten Positionen verloren.

Mit einer PSSM können DNA-Sequenzen nach Mustern durchsucht werden, die der Nukleotidhäufigkeitsverteilung der Matrix möglichst gut entsprechen (Stormo, 2000). Häufig verwendete Programme, deren Suche nach potentiellen TFBSs auf PSSMs basiert, sind MATRIX SEARCH (Prestridge und Stormo, 1993), TESS (http://www.cbil.upenn.edu/tess), MatInspector (Quandt et al., 1995) und MATCH™ (Kel et al., 2003) (siehe 3.2.2, S. 29). Allen Programmen ist gemeinsam, dass sie einen Score für die Ähnlichkeit zwischen einer bestimmten Sequenz und der PSSM berechnen. Die beiden letztgenannten Programme beziehen in die Berechnung einen sogenannten „Informationsvektor“, der Positionen der PSSM, die hoch konserviert sind, eine höhere Bedeutung für die Berechnung des Scores zukommen lässt, mit ein. Entscheidend für die Qualität der Vorhersagen sind geeignete Schwellenwerte für den berechneten Score. Im Programm MATCH™ sind für alle PSSMs in der TRANSFAC®-Datenbank (siehe 3.1.1, S. 25) vordefinierte Schwellenwerte enhalten, die daraufhin optimiert wurden, die Zahl falsch positiver, falsch negativer oder beider Arten von Vorhersagen zu minimieren.

2.2.3 Vergleichende Genomik und Phylogenetisches Footprinting

Eine weitere Methode, um funktionelle Regionen in DNA-Sequenzen vorherzusagen, ist die „vergleichende Genomik“. Sie ist ein mächtiger Ansatz zur Vorhersage von Genstrukturen und *cis*-agierenden regulatorischen Elementen (Bulyk, 2003; Cooper und Sidow, 2003; Duret und Bucher, 1997; Frazer et al., 2003; Prakash und Tompa, 2005; Ureta-Vidal et al., 2003; Xie et al., 2005). Die Vorhersage *cis*-agierender regulatorischer Elemente wird auch „phylogenetisches Footprinting“ genannt. Dieser Begriff wurde zuerst von Tagle et al. (1988) eingeführt, die γ- und ϵ-Globin-Gene in Primaten untersuchten. Die grundlegende Annahme des phylogenetischen Footprintings ist, dass regulatorische Elemente in nichtcodierenden Sequenzen während der Evolution einem höheren selektiven Druck unterliegen als nicht-funktionelle Bereiche. Schädliche Mutationen in regulatorischen Elementen werden durch Selektion aus einer Population entfernt. Daher wird angenommen, dass nicht-

funktionelle Bereiche im Laufe der Zeit mehr Mutationen ansammeln als solche, die regulatorische Elemente enthalten.

Um zu bestimmen, welche Bereiche einer Sequenz im Laufe der Evolution Mutationen angesammelt haben, vergleicht man die zu untersuchende Sequenz mit sogenannten „orthologen Sequenzen“ aus anderen Spezies. Orthologe Sequenzen sind Sequenzen, die sich aufgrund der Artenbildung (Speziation) aus einer gemeinsamen Vorläufersequenz entwickelt haben und üblicherweise ihre Funktion beibehalten. Ein Alignment (siehe 2.2.4, S. 22) orthologer, nicht-codierender Sequenzen zweier Spezies sollte daher die Bereiche, die voraussichtlich eine funktionelle Rolle haben, als konserviert hervorheben.

Aufgrund der wachsenden Verfügbarkeit vollständig sequenzierter eukaryotischer Genome wird phylogenetisches Footprinting in großem Umfang angewandt, um regulatorische Elemente und potentielle TFBSs zu identifizieren. Das inhärente Problem der vergleichenden Genomik ist allerdings die Frage, welche Spezies miteinander verglichen werden sollten, um funktionelle Bereiche möglichst zuverlässig bestimmen zu können. Einerseits sollten die zu vergleichenden Spezies nicht zu nahe verwandt sein, da sonst nicht unterschieden werden kann, ob nicht-codierende Sequenzen aufgrund evolutionären Drucks oder aufgrund der gemeinsamen Abstammung konserviert sind. Andererseits sollten die Spezies nicht zu entfernt verwandt sein, da sonst kein verlässliches Alignment nicht-codierender Sequenzen möglich ist.

Mehrere Studien haben sich mit dieser Fragestellung für *Drosophila*-Spezies beschäftigt. Bergman et al. (2002) wiesen darauf hin, dass die genomische Sequenz von *D. pseudoobscura* zur funktionellen Annotation des *D. melanogaster*-Genoms einen großen Beitrag leisten könnte. Laut dieser Studie sind Spezies wie *D. erecta* zu nahe mit *D. melanogaster* verwandt, um funktionelle Bereiche, basierend auf ihrer Konserviertheit, von nicht-funktionellen unterscheiden zu können, wohingegen das *Anopheles gambiae*-Genom keinen wesentlichen Beitrag zur Genom-Annotation leisten kann, da es im Vergleich mit *D. melanogaster* kaum Konserviertheit in nicht-codierenden Bereichen aufweist. Emberly et al. (2003) beschrieben, dass Sequenz-Konserviertheit zwischen *D. melanogaster* and *D. pseudoobscura* als alleiniges Kriterium nicht ausreicht, um Bereiche, die TFBSs enthalten, verlässlich von nicht-funktionellen Bereichen zu unterscheiden. Für 315 annotierte TFBSs fanden sie, dass deren Überlapp mit konservierten Sequenzblöcken trotz statistischer Signifikanz nicht sehr viel größer ist, als per Zufall zu erwarten ist. Es wurde jedoch gezeigt (Berman et al., 2004; Sinha et al., 2004), dass eine Kombination aus TFBS-Clusterung und vergleichender Sequenzanalyse sehr viel effektiver regulatorische Elemente in *D. melanogaster* identifiziert als jede Methode für sich alleine.

Auch für Hefe-Spezies wurde die vergleichende Genomik mit Erfolg angewandt. Für

Saccharomyces cerevisiae wurde ein Katalog regulatorischer Elemente durch eine vergleichende Analyse mit hochwertigen Rohfassungen der Genome dreier verwandter Hefe-Spezies entdeckt (Kellis et al., 2003). Diese drei Spezies besitzen eine ausreichende Sequenzähnlichkeit mit *S. cerevisiae*, um orthologe Regionen verlässlich zu alignieren, aber genug Sequenzdivergenz, um viele funktionelle Elemente aufgrund ihrer erhöhten Konserviertheit zu erkennen. Dieselben Sequenzdaten wurden auch in einer weiteren Studie verwendet (Moses et al., 2003), um die Evolution bekannter TFBSs in *S. cerevisiae* zu untersuchen. Diese TFBSs weisen weniger Substitutionen auf als der genomische Hintergrund.

Für die Identifizierung funktioneller Bereiche im menschlichen Genom haben sich Vergleiche mit dem Genom der Maus als sehr nützlich erwiesen. Mensch und Maus hatten vor ca. 75-90 Millionen Jahren einen gemeinsamen Vorfahren. Die Divergenzrate zwischen beiden Genomen war gering genug, um orthologe Sequenzen alignieren zu können, aber gleichzeitig hoch genug, um die Diskrimination funktioneller Elemente aufgrund ihrer höheren Konserviertheit zu ermöglichen. Mensch-Maus-Vergleiche wurden intensiv genutzt, um potentiell regulatorische Elemente zu identifizieren, von denen sich viele auch als funktionell herausstellten (Dermitzakis et al., 2002; Elnitski et al., 2003; Hardison et al., 1997; Loots et al., 2000).

Um die Nützlichkeit von Mensch-Maus-Vergleichen einschätzen zu können, wurde in einigen früheren Studien untersucht, in welchem Umfang experimentell bekannte TFBSs durch phylogenetisches Footprinting identifiziert werden können (Dermitzakis und Clark, 2002; Lenhard et al., 2003; Levy und Hannenhalli, 2002; Liu et al., 2004; Wasserman et al., 2000). Die Datensammlungen dieser Studien umfassten zwischen 99 und 481 TFBSs von denen 60 bis 68% durch Mensch-Maus-Vergleiche detektiert werden konnten.

2.2.4 Paarweises Sequenz-Alignment

Phylogenetisches Footprinting (siehe 2.2.3, S. 20) basiert auf dem Vergleich orthologer Sequenzen mittels paarweiser (oder auch multipler) Alignments. Im Folgenden wird kurz das Prinzip des paarweisen Sequenz-Alignments erläutert.

Beim Vergleich zweier Sequenzen stellt sich die Frage, ob diese zwei Sequenzen durch einen Prozess aus Mutation und Selektion aus einem gemeinsamen Vorfahren divergiert sind. Mutationen in DNA sind ein natürlicher evolutionärer Prozess. Fehler während der Replikation der DNA verursachen Substitutionen, d.h. es werden Nukleotide in einer Sequenz verändert, sowie Insertionen und Deletionen, d.h. es werden Nukleotide in eine Sequenz einfügt oder entfernt. Die natürliche Selektion hat einen Einfluss auf diesen Prozess,

sodass manche Mutationen häufiger als andere beobachtet werden. Um die Frage nach der Verwandtschaft zweier Sequenzen zu beantworten, werden die beiden Sequenzen gewöhnlich aligniert und dann wird entschieden, ob das paarweise Alignment auf Verwandtschaft oder Zufall basiert.

Bei einem paarweisen Alignment ordnet man die Residuen der einen Sequenz denen der anderen Sequenz so zu, dass die Reihenfolge der Residuen jeder Sequenz erhalten bleibt, und jedes Symbol der einen Sequenz einem Residuum der anderen Sequenz oder einer Lücke („gap“) zugeordnet ist (siehe Abbildung 2.7). Eine Fehlpaarung („mismatch“) im Alignment entspricht dabei einer Substitution. Lücken im Alignment weisen auf eine Deletion bzw. Insertion hin. Übereinstimmende Residuen („matches“) deuten darauf hin, dass keine Mutation stattgefunden hat.

Sequenz 1	G	A	T	A	A	A	C	T	-	T	T	T
Sequenz 2	G	-	T	A	A	T	T	T	G	T	T	T

Abbildung 2.7: Beispiel eines paarweisen Alignments. Lücken werden durch das Symbol ’-’ repräsentiert.

Die Erstellung eines Alignments ist ein Optimierungsproblem und man benötigt daher eine Scorefunktion, um erhaltene Alignments zu bewerten. Oft wird eine sogenannte „Scoringmatrix“, auch Substitutionsmatrix genannt, verwendet. Übereinstimmende Residuen deuten auf eine Verwandschaft der Sequenzen hin und sollten durch die Scoringmatrix positiv bewertet werden, wohingegen Substitutionen weniger positiv oder negativ bewertet werden sollten. Das Einfügen von Lücken wird mit einem bestimmten Wert, der sogenannten „gap penalty“, bestraft. Oft werden sogenannte „affine gap penalties“ benutzt, die eine lange Lücke weniger negativ bewerten als mehrere kurze Lücken. Im Allgemeinen wird der Score eines Alignments additiv berechnet, d.h. die Werte aus der Scoringmatrix für jedes alignierte Residuenpaar und die „gap penalties“ für jede Lücke werden aufsummiert. Dabei wird angenommen, dass Mutationen unabhängig voneinander auftreten.

Das optimale Alignment wird bestimmt, indem der Score des Alignments maximiert wird. Dabei ist zu beachten, dass das Alignment in Bezug auf das gewählte Scoringsystem optimiert wird, aber nicht zwangsweise unter biologischen Gesichtspunkten. Die meisten Algorithmen zur Bestimmung des optimalen Alignments mittels eines additiven Scoringsystems basieren auf „dynamischer Programmierung“ (Cormen et al., 2001). Es gibt zwei grundlegend verschiedene Arten des paarweisen Alignments, die auf dynamischer Programmierung beruhen, nämlich globale Alignments und lokale Alignments. Bei einem globalen Alignment werden alle Residuen beider Sequenzen einem Residuum der jeweils

anderen Sequenz oder eine Lücke zugeordnet. Im Fall eines lokalen Alignments geschieht dies nur für je eine Teilsequenz beider Sequenzen.

2.2.5 Hidden-Markov-Modelle

Hidden-Markov-Modelle (HMMs) werden in der Bioinformatik häufig zur Charakterisierung und Klassifizierung von Sequenzen eingesetzt. Ein HMM ist ein probabilistisches Modell, das durch zwei gekoppelte Zufallsprozesse beschrieben werden kann. Der erste Zufallsprozess ist eine Markov-Kette, die durch Zustände und Übergänge zwischen diesen Zuständen gekennzeichnet ist. Die einzigen Parameter, die eine Markov-Kette charakterisieren, sind die Übergangswahrscheinlichkeiten für die Übergänge zwischen den Zuständen. In einem HMM ist jeder Zustand einer Markov-Kette mit einem zweiten Zufallsprozess verknüpft, der nach einer dem Zustand zugehörigen Emissions-Wahrscheinlichkeitsverteilung eine Beobachtung emittiert. Da eine Emission nicht eindeutig einem bestimmten Zustand zugeordnet werden kann, nennt man die den Emissionen zugrunde liegenden Zustände „versteckt“ (engl. hidden). Es ist allerdings möglich, aus der beobachteten Folge von Emissionen die zugrunde liegende Folge von Zuständen abzuschätzen. Dazu bedient man sich häufig einer Methode der „dynamischen Programmierung“, nämlich des sogenannten „Viterbi-Algorithmus“ (Viterbi, 1967). Die Kenntnis der Folge von Zuständen charakterisiert die untersuchte Sequenz.

Neben vielen anderen Anwendungen in der Bioinformatik wurden HMMs auch zur Vorhersage von TFBSs eingesetzt: Z.B. entwickelten Xu et al. (2005) eine Methode, die Abhängigkeiten zwischen einzelnen Positionen von TFBSs mittels eines HMMs modelliert und exaktere Vorhersagen als PSSM-basierte Methoden liefert. Sandelin und Wasserman (2005) entwarfen ein HMM zur Vorhersage von TFBSs des TF “nuclear hormone receptor“ (NHR). NHR bindet als Dimer an zwei sogenannte „half sites“, wobei die gegenseitige Orientierung und der Abstand der beiden „half sites“ variieren kann. Das HMM modelliert dieses variable Bindeverhalten des TF. Marinescu et al. (2005) benutzten sogenannte „Profil-HMMs“, die im Gegensatz zu PSSM-basierten Methoden Insertionen und Deletionen innerhalb von TFBSs modellieren und es auch erlauben, Fragmente von TFBSs vorherzusagen.

Kapitel 3

Material und Methoden

3.1 Datenbanken

Die folgenden Abschnitte geben einen Überblick über die in dieser Arbeit benutzen Datenbanken.

3.1.1 TRANSFAC®

Die TRANSFAC®-Datenbank (Matys et al., 2003) enthält Informationen über eukaryotische TFs, deren TFBSs (siehe 2.1.3, S. 13) und die daraus abgeleiteten PSSMs (siehe 2.2.2, S. 19), die die DNA-Bindungsspezifität der TFs beschreiben. Die Daten in der TRANSFAC®-Datenbank stammen aus Veröffentlichungen, in denen Interaktionen zwischen TFs und relativ kurzen Abschnitten der DNA (Länge 5-25 bp) experimentell belegt wurden. Die Informationen aus den Veröffentlichungen werden in einem relationalen Datenmodell gespeichert. Zwischen TFs und TFBSs gibt es eine „n:m-Beziehung“, da jeder TF in der Regel an mehrere TFBSs bindet, und eine TFBS auch von unterschiedlichen TFs gebunden werden kann. Die Informationen über und die Beziehungen zwischen TFs und TFBSs werden in den Tabellen „SITES“ und „FACTORS“ der Datenbank gespeichert. In der Tabelle „SITES“ der Datenbank werden weiterhin Informationen über das von der jeweiligen TFBS regulierte Gen, die Position der TFBSs in Bezug auf dieses Gen, die untersuchte Spezies und die Methode, mit der die TFBS identifiziert wurde, gespeichert. Jeder Methode ist dabei ein bestimmter Qualitätswert von 1 bis 6 zugeordnet, der die experimentelle Verlässlichkeit einer bestimmten Protein-DNA-Interaktion widerspiegelt (siehe Tabelle 3.1).

Die Sequenzen der TFBSs werden in einer eigenen Tabelle abgelegt und wenn möglich mit einem Eintrag in der EMBL-Datenbank (Cochrane et al., 2006) verknüpft, um dem

Benutzer Zugriff auf den Kontext der TFBS-Sequenz zu ermöglichen. Die EMBL-Nukleotidsequenz-Datenbank ist Europas primäre Quelle für Nukleotidsequenzen. Die meisten DNA- und RNA-Sequenzen stammen direkt von einzelnen Forschern, aus Genomsequenzierungs-Projekten oder Patent-Anmeldungen.

Tabelle 3.1: Qualitätswerte für TFBSs in TRANSFAC®

Qualtitätswert	Beschreibung
1	Funktionell bestätigte TFBS
2	Bindung reines Proteins (gereinigt oder rekombinant) nachgewiesen
3	Immunologisch charakterisierte Bindungsaktivität eines zellulären Extraktes
4	Bindungsaktivität mittels einer bekannten Bindungssequenz nachgewiesen
5	Bindung eines uncharakterisierten Proteinauszugs an ein *bona fide*-Element
6	Keine Qualität zugewiesen

Die TRANSFAC®-Datenbank enthält in der Tabelle „MATRIX“ eine umfangreiche Sammlung von PSSMs, die basierend auf TFBSs der SITES-Tabelle konstruiert wurden. Eine PSSM ist einem oder auch mehreren TFs, deren DNA-Bindungsspezifität sich nicht erkennbar unterscheidet, zugeordnet. Einige dieser Matrizen sind nur aus TFBSs konstruiert worden, die mit Methoden, die mindestens eine bestimmte Qualität (s.o.) aufweisen, identifiziert wurden. Mit den PSSMs können genomische Sequenzen durchsucht werden, um potentielle TFBSs vorherzusagen (siehe 3.2.2, S. 29).

Es existieren zwei grundsätzliche Versionen der TRANSFAC®-Datenbank. Die sogenannte „Professional-Version“ ist kommerziell zu erwerben (http://www.biobase.de). Weiterhin gibt es eine kostenlose und frei verfügbare Public-Version, deren Datenbestand gegenüber der Professional-Version verringert ist (http://www.gene-regulation.com). Im Rahmen dieser Arbeit wurden die „Professional-Versionen“ 8.3 und 9.1 verwendet.

3.1.2 TRANSCompel®

TRANSCompel® (Kel-Margoulis et al., 2002) ist eine Datenbank mit großer Ähnlichkeit zu TRANSFAC®. Der Fokus der Datenbank liegt dabei auf sogenannten „composite elements“ (CEs). CEs bestehen aus zwei räumlich eng benachbarten TFBSs zweier TFs. Sie repräsentieren die kleinstmögliche Einheit der kombinatorischen Regulation der Transkription. Sowohl die spezifischen TF-DNA- als auch TF-TF-Interaktionen tragen zur Funktion

von CEs bei. Jeder Eintrag der Datenbank gehört zu einem bestimmten Gen und enthält Informationen über die zwei TFBSs, die zwei zugehörigen TFs und die Experimente, die die Interaktion zwischen den TFs belegen. Es gibt zwei Arten von CEs: synergistische und antagonistische: Bei synergistischen CEs führt die Interaktion zwischen den beiden TFs zu einer Erhöhung der Transkriptionsrate, die größer ist als die Summe der Transkriptionsraten, wenn beide TFs einzeln auf den Promotor wirken. In antagonistischen CEs beeinträchtigen sich die TFs gegenseitig in ihrer Wirkung.

Es existieren zwei grundsätzliche Versionen der TRANSCompel®-Datenbank. Die sogenannte „Professional-Version“ ist kommerziell zu erwerben (http://www.biobase.de). Zusätzlich gibt es eine kostenlose und frei verfügbare Public-Version, deren Datenbestand gegenüber der Professional-Version verringert ist (http://www.gene-regulation.com). Im Rahmen dieser Arbeit wurde die „Profession-Version“ 8.3 verwendet.

3.1.3 cisRED

cisRED (Robertson et al., 2006) ist eine Datenbank für konservierte regulatorische Elemente, die von einem genomweiten Vorhersagesystem identifiziert und bewertet wurden. Für ca. 7500 Sequenzsätze, die menschliche Promotoren zusammen mit durchschnittlich sechs homologen Promotoren aus anderen Vertebraten enthalten, wurden Motive mit einer Vielzahl von Methoden bestimmt. Ähnliche und überlappende Motive wurden zusammengefasst und anschließend nach statistischer Signifikanz gefiltert. Die untersuchten Sequenzen reichen von -1500 bp stromaufwärts bis 100 bp stromabwärts der jeweiligen TSS. Zur Bestimmung der Signifikanz von Motiven wurden Sequenzsätze, die aus den menschlichen Promotoren und synthetischen orthologen Promotoren bestanden, derselben Analyse unterzogen. Allen Motiven, die sowohl in den realen als auch den synthetischen Datensätzen vorkommen, wurden methodenunabhängige Scores zugeordnet. Aus der Verteilung dieser Scores für Motive in den synthetischen Datensätzen wurden p-Werte aus den Scores für Motive in den realen Datensätzen berechnet. Nur Motive mit $p < 0,05$ wurden in der Datenbank (Version 1.2e) gespeichert.

Die cisRED-Datenbank (http://www.cisred.org) ist frei verfügbar und kann lokal installiert werden.

3.1.4 Ensembl

Die Ensembl-Datenbank (Birney et al., 2006) ist eine umfassende und ganzheitliche Quelle für die Annotation von langen Genomsequenzen. Die Anzahl der verfügbaren Genome ist in den letzten Jahren von 4 auf 19 angewachsen, u.a. enthält die Datenbank die Ge-

nome von Mensch, Schimpanse, Maus, Ratte, Kuh und Hund. Die Mehrheit der Genome wird mit der automatischen Ensembl-Analyse-Pipeline (Curwen et al., 2004; Potter et al., 2004) annotiert. Der Fokus der Annotation liegt auf der Erstellung von Transkriptstrukturen. Pseudogene und RNA-Gene werden inzwischen auch von der automatischen Annotation berücksichtigt. Die Annotation von Transkriptstrukturen basiert im ersten Schritt auf Alignments von Spezies-spezifischen Protein- und cDNA-Sequenzen mit der genomischen Sequenz. In einem zweiten Schritt werden Proteine verwandter Spezies eingesetzt, um noch unentdeckte Transkripte zu lokalisieren. Die Transkripte, die auf cDNA- und Protein-Informationen basieren, werden kombiniert, um sogenannte nicht-translatierte Regionen („untranslated regions“, UTRs), der Transkripte zu ermitteln, wobei redundante Transkriptstrukturen eliminiert werden. Im letzten Schritt werden Gene basierend auf den cDNA- und Protein-basierten Transkripten annotiert.

Sammlungen von „expressed sequence tags“ (ESTs) , die bekanntermaßen viele Artefakte enthalten, werden als weniger verlässliche Quelle zur Vorhersage von Genstrukturen eingeschätzt und zu einer separaten Vorhersage von sogenannten „EST-Genen“ genutzt. Diese EST-Sammlungen wurde aufgrund des Fehlens anderer Daten für die Genome von Huhn und Honigbiene zur Annotation der Gene, die normalerweise basierend auf Protein- und cDNA-Sequenzen annotiert werden, herangezogen. Die EST-Sammlungen für diese Spezies wurden von wenigen Gruppen erstellt und sollten daher von konsistenter Qualität sein.

Seit dem „Release 34“ enthält die Ensembl-Datenbank auch Informationen über regulatorische Elemente. Dazu wurden externe Quellen integriert, die genomweite Vorhersagen regulatorischer Elemente enthalten. Die ersten integrierten Datensätze sind die cisRED-Datenbank (siehe 3.1.3, S. 27) und die mittels MiRanda (John et al., 2004) vorhergesagten Zielsequenzen von microRNAs.

Die Ensembl-Datenbank (http://www.ensembl.org) ist frei verfügbar und kann lokal installiert werden. Ein Zugriff ist auch über die bereitgestellten Perl- und JAVA-APIs möglich (Stabenau et al., 2004). Für MySQL-Clients ist sie zusätzlich über das Internet verfügbar (Host-Adresse: ensembldb.ensembl.org).

3.2 Programme

Im Folgenden wird ein Überblick über die im Rahmen dieser Arbeit verwendeten Programme gegeben.

3.2.1 RepeatMasker

RepeatMasker (http://www.repeatmasker.org) ist ein Programm, das DNA-Sequenzen nach sich wiederholenden Abschnitten, sogenannten „Repeats“, durchsucht. Das Programm verlangt als Eingabe eine Anfragesequenz. RepeatMasker gibt sowohl eine detaillierte Annotation der Repeats in der Anfragesequenz als auch eine modifizierte Version der Anfragesequenz aus, in der alle Repeats maskiert sind, wobei diese im Normalfall durch ’N’s ersetzt werden. Durchschnittlich werden ca. 50% der genomischen Sequenz des Menschen durch das Programm maskiert.

3.2.2 MATCH™

MATCH™ ist ein Programm, das PSSMs (siehe 2.2.2, S. 19) benutzt, um in DNA-Sequenzen nach potentiellen TFBSs zu suchen. Eine umfangreiche Sammlung dieser PSSMs ist in der TRANSFAC®-Datenbank (siehe 3.1.1, S. 25) enthalten. MATCH™ berechnet die Ähnlichkeit zwischen einer Sequenz und einer PSSM mit zwei Werten: dem sogenannten „matrix similarity score“ (MSS) und dem „core similarity score“ (CSS) . Diese Scores können Werte zwischen 0 und 1 annehmen, wobei ein Wert von 1 für eine maximale Ähnlichkeit der untersuchten Sequenz zur PSSM steht.

Der Kern („core“) jeder PSSM sind die fünf am stärksten konservierten, benachbarten Nukleotide. Die Berechnung des CSS dient der Beschleunigung des Algorithmus zur Berechnung des MSS: Der zugrunde liegende Gedanke ist, den MSS nur dann zu berechnen, wenn der CSS einen bestimmten Schwellenwert überschreitet, d.h. erst wenn eine genügend große Ähnlichkeit im Kern der PSSM vorliegt, wird die gesamte Ähnlichkeit der untersuchten Sequenz zur PSSM berechnet. Der Algorithmus gibt nur die Treffer der PSSM aus, für die beide Scores höher als die entsprechenden Schwellenwerte sind.

Der MSS (wie auch der CSS) für eine Sequenz x der Länge L wird folgendermaßen berechnet:

$$\text{MSS} = \frac{\sum_{i=1}^{L} I(i) \cdot f_{i,b_i} - \sum_{i=1}^{L} I(i) \cdot f_i^{min}}{\sum_{i=1}^{L} I(i) \cdot f_i^{max} - \sum_{i=1}^{L} I(i) \cdot f_i^{min}} \tag{3.1}$$

$$\text{mit} \quad I(i) = \sum_{b \in \{A,C,G,T\}} f_{i,b} \cdot \ln\left(\frac{f_{i,b}}{0.25}\right) \tag{3.2}$$

wobei f_{i,b_i} die relative Häufigkeit des Nukleotids b an Position i der Matrix ($b \in \{A, C, G, T\}$), f_i^{min} die relative Häufigkeit des seltensten und f_i^{max} die relative Häufigkeit des häufigsten Nukleotids an Position i der Matrix ist.

Der Informationsvektor $I\,(i)$ beschreibt die Konserviertheit der Positionen der PSSM (Quandt et al., 1995). Durch die Multiplikation der relativen Häufigkeiten mit dem Informationsvektor werden Mismatches in weniger konservierten Positionen eher akzeptiert, und gleichzeitig in konservierten Regionen stark benachteiligt. Kel et al. (1999) zeigten, dass die Verwendung des Informationsvektors im Vergleich mit Methoden, die diesen nicht nutzen, zu einer besseren Performanz bei der Erkennung von TFBSs führt.

3.2.3 Alignment-Programme

Ein sehr wichtiges Werkzeug in der Sequenzanalyse ist das Alignieren von Sequenzen und viele Algorithmen wurden für das Problem des Sequenzalignments entwickelt (Batzoglou, 2005; Notredame, 2002). Zwei grundlegende Arten des Alignments sind globale und lokale Alignments, die im Folgenden genauer beschrieben werden.

Globales Alignment

Bei einem globalen Alignment werden zwei Sequenzen $x = x_1 \ldots x_m$ und $y = y_1 \ldots y_n$ über ihre gesamte Länge aligniert. Der erste Ansatz zum globalen Alignment zweier Sequenzen wurde von Needleman und Wunsch (1970) eingeführt. Eine verbesserte Version dieses Algorithmus wurde von Gotoh (1982) vorgeschlagen. Die zugrunde liegende Idee ist, ein optimales Alignment aus vorher berechneten Lösungen von optimalen Alignments kürzerer Teilsequenzen zu bilden.

Dazu wird eine Matrix F konstruiert, die mit einem Index i für die Sequenz x und einem Index j für die Sequenz y indiziert wird. Der Wert $F(i, j)$ ist dabei der Score des optimalen Alignments der Präfixe $x_1 \ldots x_i$ und $y_1 \ldots y_j$. Die Matrix $F(i, j)$ wird rekursiv gefüllt. Dazu benötigt man eine Scoringmatrix s, die jedem Paar von Residuen (x_i, y_j) einen bestimmten Score $s(x_i, y_j)$ zuordnet. Die Kosten eine Lücke einzufügen, sind durch $d > 0$ gegeben. Die Zeile $F(0, j)$ bzw. Spalte $F(i, 0)$ der Matrix entspricht dabei der Position vor dem ersten Residuum der Sequenz x bzw. y. Die Matrix F wird mit $F(0, 0) = 0$ initialisiert. Die Werte $F(i, 0)$ sind die Scores eines Alignments des Präfix $x_1 \ldots x_i$ mit Lücken in Sequenz y und können daher mit $F(i, 0) = -id$ initialisiert werden. Entsprechend wird $F(0, j) = -jd$ definiert.

Die Matrix wird von der linken oberen Ecke ausgehend gefüllt. Es gibt drei Möglichkeiten den besten Score für ein Alignment der Präfixe $x_1 \ldots x_i$ und $y_1 \ldots y_j$ zu erhalten: x_i

kann zu y_j aligniert werden, d.h. $F(i,j) = F(i-1,j-1) + s(x_i, y_i)$; oder x_i ist zu einer Lücke aligniert, d.h. $F(i,j) = F(i-1,j) - d$; oder y_i ist zu einer Lücke aligniert, d.h. $F(i,j) = F(i,j-1) - d$. Wenn $F(i-1,j-1)$, $F(i-1,j)$ und $F(i,j-1)$ bekannt sind, wird $F(i,j)$ als das Maximum dieser drei Möglichkeiten berechnet:

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_i) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases} \tag{3.3}$$

Für jeden berechneten Wert $F(i,j)$ wird ein Zeiger auf die Zelle der Matrix, aus der $F(i,j)$ berechnet wurde, gerichtet. Der Wert $F(m,n)$ ist der optimale Score eines Alignments von $x_1 \ldots x_m$ zu $y_1 \ldots y_n$. Um das Alignment selbst zu erhalten, müssen die Zeiger von diesem finalen Wert bis zum Eintrag $F(0,0)$ zurückverfolgt werden („traceback“).

Lokales Alignment

In manchen Fällen sucht man nicht nach dem besten globalen Alignment zweier Sequenzen, sondern nach dem optimalen Alignment zwischen Teilsequenzen von x und y. Dies trifft zum Beispiel auf Proteinsequenzen, bei denen erwartet wird, dass sie eine gemeinsame Domäne aber sonst nur eine geringe Ähnlichkeit aufweisen, oder lange genomische Sequenzen zu. Vergleicht man zwei stark divergente Sequenzen, so ist ein lokales Alignment die sensitivste Methode, eine Ähnlichkeit zwischen diesen festzustellen. Das Alignment zweier Teilsequenzen von x und y mit dem höchsten Score wird als das optimale lokale Alignment von x und y bezeichnet.

Der Algorithmus zur Bestimmung dieses optimalen lokalen Alignments wurde von Smith und Waterman (1981) eingeführt und basiert auf dem Algorithmus zur Bestimmung des optimalen globalen Alignments von Needleman und Wunsch (1970). Es gibt zwei Unterschiede: Erstens kann jede Zelle der Matrix $F(i,j)$ zusätzlich zu den Möglichkeiten in Gleichung (3.3) den Wert 0 annehmen, wenn diese alle Werte kleiner 0 besitzen:

$$F(i,j) = \max \begin{cases} 0 \\ F(i-1,j-1) + s(x_i, y_i) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases} \tag{3.4}$$

Die Wahl des Wertes 0 entspricht dem Beginn eines neuen Alignments. Wenn das beste Alignment bis zu einem gewissen Punkt einen negativen Score hat, ist es besser, ein neues

Alignment zu starten, als ein altes weiter zu verlängern. Eine Konsequenz daraus ist, dass die Zellen der ersten Zeile bzw. Spalte der Matrix jeweils den Wert $F(0, j) = 0$ bzw. $F(i, 0) = 0$ annehmen.

Der zweite Unterschied ist, dass das Alignment nun in jeder Zelle der Matrix enden und beginnen kann, d.h. der höchste Score des Alignments wird nicht mehr der Zelle $F(m, n)$ entnommen, sondern der Zelle $F(i, j)$ mit dem höchsten Score in der gesamten Matrix. Ausgehend von dieser Zelle werden die Zeiger bis zu einer Zelle, die den Wert 0 annimmt, zurückverfolgt.

Ein kurzer Überblick über die in dieser Arbeit benutzen Alignment-Programme wird im Folgenden gegeben.

3.2.3.1 WU-BLAST

Das sogenannte „Basic Local Alignment Search Tool“ (BLAST) (Altschul et al., 1990) dient dem Finden von lokalen Ähnlichkeiten zwischen Sequenzen. Dazu werden Protein- oder Nukleotidsequenzen mit Sequenzen in einer Datenbank verglichen und die statistische Signifikanz von Ähnlichkeiten berechnet. Die Entdeckung von Homologien zu bekannten Sequenzen kann dabei erste Hinweise auf die Funktion einer unbekannten Sequenz geben. BLAST ist eine heuristische Methode und daher im Vergleich mit Algorithmen, die alleinig auf dynamischer Programmierung basieren, sehr schnell.

Die BLAST zugrunde liegende Idee ist, dass korrekte Alignments höchstwahrscheinlich einen kurzen Bereich von (nahezu) identischen Residuen enthalten. In einem ersten Schritt wird daher nach solchen Bereichen gesucht, um die herum das Alignment in einem zweiten Schritt erweitert wird. Aus der Anfragesequenz wird dazu eine Liste kurzer sogenannter „seeds“ erstellt, deren Länge für Proteine drei Aminosäuren und für DNA elf Nukleotide beträgt. BLAST erstellt dann eine Liste aller möglichen Wörter, die mit einem der „seeds“ eine Ähnlichkeit oberhalb eines gewissen Schwellenwertes T aufweisen. Die Datenbank wird nach den Wörtern aus dieser Liste durchsucht und sobald ein Treffer gefunden wird, startet die sogenannte „ungapped extension“, d.h. das Alignment wird in beide Richtungen ohne Lücken erweitert, bis der maximale Score erreicht wird. Dieses Alignment nennt man ein „high-scoring segment pair“ (HSP).

WU-BLAST ist eine Weiterentwicklung des ursprünglichen BLAST-Programms (Altschul et al., 1990) und basiert auf dem „gapped BLAST“-Algorithmus (Altschul et al., 1997). Der ursprüngliche BLAST-Algorithmus behandelt Lücken nur implizit, da in vielen Fällen mehrere verschiedene HSPs auf derselben Datenbanksequenz lokalisiert sind und die statistische Signifikanz ihrer Kombination berechnet wird. Wird allerdings eines dieser HSPs übersehen, kann die statistische Signifikanz ihrer Kombination verloren gehen.

Durch den „gapped BLAST“-Algorithmus muss nur noch ein Alignment statt aller lückenlosen Alignments, die in einem signifikanten Ergebnis zusammengefasst sind, gefunden werden, d.h. es reicht, nur ein zugehöriges HSP zu finden, um das kombinierte Ergebnis erhalten zu können. Liegen zwei nicht-überlappende Treffer mit Score $\geq T$ innerhalb eines maximalen Abstandes A voneinander, wird der zweite Treffer wie oben beschrieben mittels der „ungapped extension“ verlängert . Besitzt das so erzeugte HSP einen Score, der einen bestimmten moderaten Score S_g überschreitet, wird es mittels dynamischer Programmierung in beide Richtungen verlängert. Dieser Schritt wird „gapped extension“ genannt. Der Score S_g wird dabei so gewählt, dass die „gapped extension“ im statistischen Mittel nur für eine von 50 Datenbanksequenzen durchgeführt wird. Unterschreitet das erhaltene Alignment einen Erwartungswert E, der die zufällig erwartete Anzahl von Alignments mit gleichem Score widergibt, wird es für den Benutzer ausgegeben.

3.2.3.2 BLASTZ

Das Programm BLASTZ (Schwartz et al., 2003) ist eine unabhängige Implementation des „Gapped BLAST“-Algorithmus (Altschul et al., 1997), die auf drei Schritten zur Erstellung des Alignments beruht:

(1) Dem Auffinden kurzer fast identischer Sequenz-Abschnitte, (2) dem Verlängern dieser Abschnitte, ohne dabei Lücken zu erlauben, und (3) dem Verlängern jedes lückenfreien Abschnittes, der einen gewissen Schwellenwert überschreitet, mittels dynamischer Programmierung, wobei nun das Einfügen von Lücken erlaubt ist.

Im Unterschied zu „Gapped BLAST“ besteht bei BLASTZ erstens die Möglichkeit, zu verlangen, dass übereinstimmende Regionen in beiden Sequenzen in gleicher Orientierung und Reihenfolge vorkommen, und zweitens benutzt BLASTZ ein Alignment-Scoring-Schema, das von Chiaromonte et al. (2002) entworfen und getestet wurde. Dieses erschwert es, dass Bereiche mit extrem verzerrtem Nukleotidgehalt den obigen Schritt (3) auslösen.

3.2.3.3 BLAT

BLAT (Kent et al., 2002) ist das sogenannte „BLAST-like alignment tool“ zum schnellen Alignieren von DNA-Sequenzen. Ein Haupteinsatzgebiet von BLAT ist die Mappierung von mRNA-Sequenzen auf ihre entsprechenden genomischen Sequenzen. BLAT ist am effizientesten, wenn die zu alignierenden Sequenzen eine Identität $\geq 90\%$ aufweisen. In vielen Punkten ähnelt es BLAST: Das Programm sucht nach kurzen, gut übereinstimmenden Abschnitten, sogenannten „hits“, und verlängert diese in HSPs. In einigen Punkten unterscheidet sich BLAT jedoch erheblich von BLAST: BLAST bildet einen Index der

Anfragesequenz und durchsucht dann linear die Datenbank. BLAT geht den umgekehrten Weg, d.h. es bildet einen Index der Datenbank und durchsucht dann linear die Anfragesequenz. Die Bildung des Index erfordert einmalig eine hohe Rechenzeit, aber dieser Index kann für mehrere Anfragen benutzt werden. Dies bedeutet einen immensen Geschwindigkeitsvorteil, da nur noch die relativ kurzen Anfragesequenzen, die meist eine Länge von einigen kb haben, durchsucht werden müssen und nicht die gesamte Datenbank, die für das menschliche Genom aus ca. 3.2 Gb besteht. Weiterhin gibt BLAST jeden homologen Bereich zwischen zwei Sequenzen einzeln zurück, wohingegen BLAT diese zu einem einzigen Alignment verknüpft.

3.2.3.4 AVID

AVID (Bray et al., 2003) ist ein globales Alignment-Programm, das entworfen wurde, um lange genomische Sequenzen, bei denen der Ansatz der dynamischen Programmierung aufgrund deren Länge versagt, alignieren zu können. AVID zerlegt Sequenzen rekursiv in kleinere Sequenzen, die schließlich mit dem Needleman-Wunsch-Algorithmus aligniert werden können.

Diese Zerlegung geschieht mittels lokaler Verankerung, d.h. es werden zunächst lange Abschnitte gesucht, die in beiden Sequenzen exakt übereinstimmen und deren flankierende Regionen eine hohe Ähnlichkeit aufweisen. Diese Abschnitte werden Anker genannt. Sie werden Teil des finalen Alignments und nicht mehr verändert. Der beschriebene Vorgang wird für die Regionen zwischen den Ankern rekursiv wiederholt, wobei sukzessiv kleinere Abschnitte exakter Übereinstimmung zur weiteren Ankerbestimmung benutzt werden. Schließlich werden die verbleibenden kurzen Regionen zwischen den Ankern mit klassischer dynamischer Programmierung aligniert.

3.2.3.5 CLUSTALW

CLUSTALW (Thompson et al., 1994) berechnet Alignments auf der Grundlage dynamischer Programmierung, unterschiedet sich aber vom klassischen Needleman-Wunsch-Algorithmus dadurch, dass Strafen für Lücken, die sogenannten „gap penalties“, abhängig von der Position und des zu alignierenden Restes variiert werden und auch die geschätzte Divergenz der zu alignierenden Sequenzen berücksichtigt wird.

CLUSTALW ist ein Programm, das häufig für die Erzeugung multipler Alignments angewandt wird. Diese werden von CLUSTALW in einer progressiven Art und Weise erzeugt, wie sie von Feng und Doolittle (1987) beschrieben wurde. Dazu werden alle paarweisen Alignments gebildet, um eine Distanzmatrix zu berechnen, die die paarweise Diver-

genz aller Sequenzpaare enthält. Das multiple Alignment wird progressiv gebildet, indem die Sequenzen entsprechend der Verzweigungsabfolge eines sogenannten „guide tree“ aligniert werden. Der „guide tree“ wird nach der Neighbour-Joining Methode (Saitou und Nei, 1987) aus der o.g. Distanzmatrix berechnet. Diese Methode produziert Bäume deren Kantenlängen proportional zur geschätzten Divergenz entlang jeder Kante sind. Das Hinzufügen von Sequenzen zum multiplen Alignment wird mit dem Needleman-Wunsch-Algorithmus durchgeführt.

3.2.3.6 CONREAL

Im Gegensatz zu anderen Alignment-Programmen ist CONREAL (Berezikov et al., 2004) speziell zum Alignieren von nicht-codierenden regulatorischen Sequenzen entwickelt worden. Zum Alignieren von codierenden Sequenzen ist es hingegen nicht geeignet. CONREAL aligniert orthologe Sequenzen, indem es vorhergesagte konservierte regulatorische Elemente zur Verankerung benutzt. Dabei wird angenommen, dass die Sequenz und Anordnung regulatorischer Elemente zwischen orthologen Promotoren größtenteils konserviert ist.

CONREAL durchsucht orthologe Promotoren mit einer Sammlung von PSSMs, z.B. alle PSSMs für Mensch, Maus und Ratte aus TRANSFAC®, nach TFBSs, wobei für jede PSSM Start, Ende, Strang und Score der Treffer gespeichert werden. Für eine gegebene PSSM werden nun alle vorhergesagten TFBSs der ersten orthologen Sequenz mit allen Treffern, die auf dem gleichen Strang in der zweiten orthologen Sequenz liegen, verglichen. Für die Treffer und vordefinierte flankierende Sequenzen einer Länge von 5 bis 20 bp wird die prozentuelle Sequenzidentität berechnet. Alle paarweisen Vergleiche, die eine bestimmte prozentuelle Sequenzidentität zwischen 10 und 50% unterschreiten, werden verworfen. Die Liste der verbleibenden paarweisen Vergleiche wird im Anschluss daran zuerst nach absteigender Sequenzidentität und danach nach absteigendem Informationsgehalt der PSSMs sortiert. Diese sortierte Liste wird von oben nach unten durchlaufen, wobei Paare verankert und damit aligniert werden, die keinen vorherigen Paaren widersprechen, die schon als Anker fungieren. Bereiche, die zwischen den Ankern liegen, d.h. in denen keine konservierten TFBSs vorhergesagt wurden, werden allerdings nicht aligniert. CONREAL produziert daher kein globales Alignment im klassischen Sinne, sondern eine geordnete Abfolge von konservierten vorhergesagten TFBSs. Durch die Matrizensuche ist CONREAL sehr rechenintensiv und benötigt längere Laufzeiten als alle anderen in dieser Arbeit eingesetzten Alignment-Programme.

3.2.3.7 DIALIGN

DIALIGN (Morgenstern et al., 1998; Morgenstern, 1999) verfolgt einen segment-basierten Ansatz und unterscheidet sich daher grundlegend vom Needleman-Wunsch-Algorithmus. Das Alignment zweier Sequenzen wird durch den Vergleich ganzer Segmente, d.h. lückenloser Abschnitte von Residuen, anstatt einzelner Residuen der Sequenzen gebildet.

Die Motivation für diesen Ansatz sind einige Eigenschaften des Needleman-Wunsch-Algorithmus. Dieser produziert vernünftige Alignments, wenn die zu alignierenden Sequenzen eine hohe Ähnlichkeit aufweisen und nur eine kleine Anzahl von Lücken eingefügt werden muss. Allerdings hat er auch einige wohlbekannte Schwächen: Ähnliche Regionen in Sequenzen, die insgesamt eine geringe Ähnlichkeit aufweisen, werden kaum aligniert, und benutzerdefinierte Parameter, wie z.B. die „gap penalty“, beeinflussen das erhaltene Alignment erheblich.

DIALIGN legt einige dieser Schwächen ab, indem es keine „gap penalty“ verwendet und sich nicht auf den Vergleich einzelner Residuen konzentriert. Bereiche geringer Ähnlichkeit werden vom Alignment ausgeschlossen, was im Gegensatz zum Needleman-Wunsch-Algorithmus die Detektion und das Alignieren kurzer ähnlicher Regionen in Sequenzen von insgesamt geringer Ähnlichkeit ermöglicht. Die zu vergleichenden Segmente haben somit die gleiche Länge und dürfen keine Lücken enthalten. Paare von Segmenten werden als Diagonalen bezeichnet, da sie in der Alignment-Matrix beider Sequenzen, die auch „dot plot“ oder „dot matrix“ (Sonnhammer und Durbin, 1995) genannt wird, als Diagonalen erscheinen würden. Eine Gruppe solcher Diagonalen wird als „konsistent“ bezeichnet, wenn es keine doppelten oder sich überkreuzenden Zuordnungen zwischen den Residuen beider Sequenzen gibt. Anschließend werden die Diagonalen jeder konsistenten Gruppe unter probabilistischen Gesichtspunkten gewichtet. Aus der konsistenten Gruppe von Diagonalen, für die die Summe der Gewichte der Diagonalen maximal wird, wird das optimale Alignment konstruiert. Eine detaillierte Beschreibung des Algorithmus findet sich in Morgenstern et al. (1996).

Für die Berechnung eines multiplen Alignments identifiziert DIALIGN zuerst stark homologe Segmentpaare, die mit einem p-Wert ähnlich dem E-Wert bei BLAST gewichtet werden. Jedes Segmentpaar wird zusätzlich mit einem weiteren Score, dem sogenannten „overlapping weight“, versehen, der proportional zur Kompatibilität dieses Segmentpaares mit dem kompletten Satz von Segmentpaaren ist. DIALIGN setzt dann das Alignment progressiv, aber in einer Sequenz-unabhängigen Weise, zusammen. Dazu werden Segmentpaare der Reihe nach entsprechend ihrer Scores, beginnend mit dem höchsten Score, kombiniert, bis alle Residuen aus jeder Sequenz in das multiple Alignment eingefügt worden sind.

3.2.3.8 LAGAN und Multi-LAGAN

LAGAN und Multi-LAGAN wurden als „effiziente Werkzeuge zum multiplen Alignieren von genomischer DNA im großen Maßstab“ (Brudno et al., 2003) entwickelt.

LAGAN steht für „Limited Area Global Alignment of Nucleotides“ und nutzt wie andere effiziente globale Alignment-Algorithmen, wie z.B. AVID, die Technik der Verankerung. Das Alignieren zweier Sequenzen verläuft in drei Stufen: (1) dem Erzeugen lokaler Alignments beider Sequenzen, (2) der Begrenzung eines Gebietes, „rough global map“ genannt, in der Alignment-Matrix durch die Aneinanderkettung einer geordneten Teilmenge der lokalen Alignments, und (3) der Berechnung des besten globalen Alignments, das innerhalb der „rough global map“ in der Alignment-Matrix liegt.

Effiziente globale Alignment-Algorithmen hängen von der Verankerung ab und benötigen daher eine sensitive Methode zum Erkennen lokaler Alignments. AVID benutzt exakt übereinstimmende Abschnitte als Anker, was bei insgesamt sehr ähnlichen Sequenzen, wie z.B. orthologen Mensch- und Maus-Sequenzen, gut funktioniert. LAGAN hingegen verwendet einen Ansatz zur Verankerung, der sowohl für nah als auch entfernt verwandte Organismen ausgelegt wurde. Anker werden dabei mit dem CHAOS-Algorithmus (Brudno und Morgenstern, 2002) vorhergesagt. CHAOS ist eine hochsensitive Methode zur Bestimmung lokaler Alignments, bei der statt langer und exakter Abschnitte kurze und ungenaue verwendet werden. LAGAN berechnet das globale Alignment, indem zuerst CHAOS rekursiv auf Sequenzabschnitte mit wenig Ankern angewendet wird, bis jedes aufeinanderfolgende Paar von Ankern einen gewissen Abstand unterschreitet. Anschließend wird das finale Alignment mittels dynamischer Programmierung auf dem begrenzten Gebiet der Alignment-Matrix, d.h. in der „rough global map“, um und zwischen den Ankern, gebildet.

Ein multiples Alignment wird von Multi-LAGAN progressiv gebildet, indem nacheinander zwei Sequenzen, oder multiple Alignments aus Zwischenschritten, mit dem LAGAN-Algorithmus aligniert werden. In einer optionalen iterativen Verbesserungsphase wird jede Sequenz aus dem Alignment entfernt und zum Rest des Alignments re-aligniert, bis keine weiteren Verbesserungen mehr gefunden werden. Im Gegensatz zu CLUSTALW berechnet Multi-LAGAN keinen „guide tree“ basierend auf einer Distanzmatrix der einzelnen Sequenzen, sondern benötigt einen solchen phylogenetischen Baum als Eingabe.

3.2.3.9 LALIGN

LALIGN (Huang und Miller, 1991) ist eine effizientere Version des Waterman-Eggert-Algorithmus (Waterman und Eggert, 1987). Dieser ist die häufigst genutzte Methode, um verschiedene suboptimale paarweise Alignments zu berechnen. Dieser Algorithmus findet

das jeweils beste Alignment, das keine alignierten Residuenpaare mit dem zuvor erhaltenen Alignment gemeinsam hat. Zuerst wird der Smith-Waterman-Algorithmus angewandt, um das beste lokale Alignment zweier Sequenzen zu berechnen. Im nächsten Schritt werden alle Zellen der Alignment-Matrix, die mit Residuenpaaren dieses besten lokalen Alignments korrespondieren, eliminiert, damit sie nicht zur Berechnung des nächsten Alignments beitragen können. Aus der resultierenden Alignment-Matrix wird der Score für das nun beste lokale Alignment berechnet. Dieser Vorgang kann wiederholt werden: LALIGN berechnet standardmäßig die besten zehn nicht-überlappenden lokalen Alignments zweier Sequenzen.

3.2.3.10 T-COFFEE

T-COFFEE wurde als „neuartige Methode für schnelle und akkurate multiple Sequenzalignments“ (Notredame et al., 2000) entwickelt. T-COFFEE aligniert Sequenzen in einer progressiven Weise, benutzt dabei aber eine konsistenzbasierte Zielfunktion namens COFFEE („Consistency based Objective Function For alignmEnt Evaluation“) (Notredame et al., 1998). Die zugrundliegende Idee von COFFEE ist, dass, wenn zwei Sequenzen a und b während eines progressiven Alignment-Schrittes aligniert werden, später nicht mehr zu korrigierende Fehler vermieden werden können, indem vorzugsweise Positionen aligniert werden, die konsistent zu den gleichen Positionen einer dritten Sequenz c in nachfolgenden Alignment-Schritten aligniert werden.

T-COFFEE bildet erst alle globalen und lokalen paarweisen Alignments der Eingabesequenzen mit CLUSTALW bzw. LALIGN. Den paarweisen Alignments werden Gewichte entsprechend ihrer Sequenzidentität zugewiesen. Jedes paarweise Alignment wird dann als eine Liste gewichteter, paarweiser Übereinstimmungen pro Residuum repräsentiert, wie z.B. Nukleotid X aus Sequenz a ist zu Nukleotid Y aus Sequenz b aligniert. Wenn ein beliebiges Paar von Residuen sowohl im globalen als auch im lokalem Alignment vorkommt, wird es in einem einzigen Eintrag, dessen Gewicht der Summe der beiden einzelnen Gewichte entspricht, in der Liste zusammengefasst. In einem nachfolgenden Schritt werden diese Gewichte so modifiziert, dass sie widerspiegeln, inwiefern diese Reste konsistent mit anderen Resten aus allen anderen Sequenzen alignierbar sind. Ähnlich zu CLUSTALW wird ein „neighbor joining“-Baum aus allen paarweisen Distanzen der Sequenzen abgeschätzt, und die Sequenzen werden nacheinander, der Topologie des Baumes folgend, aligniert. Die zwei benachbartesten Sequenzen werden mittels dynamischer Programmierung unter Verwendung der beschriebenen Gewichte aligniert. Dieses Alignment ist fixiert, d.h. Lücken darin können nicht mehr verschoben werden. Der Hauptunterschied zu CLUSTALW ist der, dass die Information aus allen paarweisen Alignments während jedes pro-

gressiven Alignment-Schrittes genutzt wird und nicht nur die Information der Sequenzen, die während des aktuellen Schrittes aligniert werden.

3.3 Daten und Datenanalyse

3.3.1 Datensätze

Im Rahmen dieser Arbeit wurden vier verschiedene Datensätze von TFBSs auf ihre Konserviertheit hin untersucht. Die Daten entstammen dabei den Datenbanken TRANSFAC®, TRANSCompel® und cisRED (siehe 3.1, S. 25).

Datensatz I besteht aus 2676 TFBSs (Mensch: 1349, Maus: 812, Ratte: 515), die aus der TRANSFAC®-Datenbank (Professional Release 9.1, siehe 3.1.1, S. 25) extrahiert wurden. Diese TFBSs wurden im Vergleich mit entsprechenden orthologen Maus- bzw. Mensch-Sequenzen untersucht.

Datensatz II besteht aus 20322 sogenannten „Regulatory Features“ (RFs) , die der cisRED-Datenbank (siehe 3.1.3, S. 27) entstammen und für Promotoren auf dem Chromosom 1 des Menschen vorhergesagt wurden. Diese RFs wurden mit orthologen Maus-Sequenzen verglichen.

Datensatz III besteht aus 405 TFBSs (Mensch: 241, Maus: 108, Ratte: 56), die in CEs vorkommen und aus der TRANSCompel®-Datenbank (Professional Release 8.3, siehe 3.1.2, S. 26) extrahiert wurden. Diese wurden im Vergleich mit entsprechenden orthologen Maus- bzw. Mensch-Sequenzen auf ihre Konserviertheit hin untersucht.

Datensatz IV besteht aus 928 Mensch-TFBSs, die aus der TRANSFAC®-Datenbank (Professional Release 9.3) extrahiert wurden und für die orthologe Sequenzen in Maus, Ratte, Kuh und Hund gefunden wurden.

3.3.2 Mappierung von Transkriptionsfaktor-Bindestellen

Für alle TFBSs in TRANSFAC® und TRANSCompel® wurde experimentell bestätigt, dass ein TF an sie bindet. Um die genomischen Lokalisationen dieser TFBSs zu erhalten, wurden sie auf ihre entsprechenden Genome mappiert. Dies wird im Folgenden beschrieben:

Die Zeichenketten der TFBS-Sequenzen in den Datenbanken unterscheiden sich dadurch, dass ein Anteil Groß- und Kleinbuchstaben enthält, der Rest nur Großbuchstaben. Für TFBS-Sequenzen, die aus Groß- und Kleinbuchstaben bestehen, wurden nur die Großbuchstaben einer TFBS-Sequenz als die eigentliche TFBS angesehen. Die TFBSs wurden mittels ihrer Verweise auf die EMBL-Datenbank, die in TRANSFAC® angegeben sind, um 50 bp auf jeder Seite verlängert, um eine eindeutige Mappierung auf das korrespondierende Genom (siehe Tabelle 3.2) mittels BLAT (siehe 3.2.3.3, S. 33) zu ermöglichen. Entsprechend der erhaltenen genomischen Koordinaten wurden die TFBSs mit einer 800 bp

Umgebung versehen.

Die mappierten TFBSs wurden mittels der Genom-Annotation in der Ensembl-Datenbank (siehe 3.1.4, S. 27) Genen zugeordnet. Dieser Schritt war nötig, da ein Teil der TFBSs in TRANSFAC® nicht mit einem Ensembl-Gen verknüpft ist. Da die TFBS-Sequenzen strang-spezifisch annotiert sind, wurde jede TFBS demjenigen Ensembl-Gen zugeordnet, das auf dem gleichen Strang wie die TFBS liegt und das Transkript mit dem geringsten Abstand zwischen seiner TSS und der TFBS aufweist. Die benutzten Ensembl-Versionen sind Tabelle 3.2 zu entnehmen.

Tabelle 3.2: Zur Analyse verwendete Ensembl- und Genom-Versionen

Datensatz	Ensembl	Mensch	Maus	Ratte	Hund	Kuh
I	30	build 35	build m33	RGSC 3.4	-	-
II	34	build 35	build m33	-	-	-
III	27	build 35	build m33	RGSC 3.1	-	-
IV	33	build 35	build m33	RGSC 3.4	CanFam1.0	Btau 1.0

3.3.3 Beschaffung orthologer Sequenzen

Konnte eine TFBS einem Gen zugeordnet werden, wurde die Ensembl-Datenbank ensembl_mart (siehe Tabelle 3.2) nach bekannten orthologen Genen durchsucht, d.h. für die Datensätze I bis III nach orthologen Maus-Genen für TFBSs des Menschen und orthologen Mensch-Genen für TFBS der Maus und der Ratte. Für Datensatz IV wurde die ensembl_mart-Datenbank nach orthologen Maus-, Ratte-, Hund- und Kuh-Genen durchsucht.

Wenn ein orthologes Gen existierte, wurden die erhaltenen Sequenzen mit WU-BLAST (Gish,W., http://blast.wustl.edu) gegen das zweite Genom aligniert. Die jeweils verwendete Genom-Version ist Tabelle 3.2 zu entnehmen. Eine Suche gegen das komplette Genom wurde gegenüber einer Suche, die auf Regionen nahe der bekannten orthologen Gene eingeschränkt ist, bevorzugt, um eine willkürliche Wahl der Größe dieser Regionen zu vermeiden. Vor der WU-BLAST-Suche wurden potentielle Repeats mit dem Programm RepeatMasker (siehe 3.2.1, S. 29) in Kombination mit MaskerAid (Bedell et al., 2000) und der Repeat-Datenbank „RepBaseUpdate“ (März 2004) (Jurka, 2000) Spezies-spezifisch maskiert. Die Sequenz um den WU-BLAST-Treffer, der den höchstmöglichen Score hat und gleichzeitig einem bekannten orthologen Gen zugeordnet wurde (siehe 3.3.2, S. 40), wurde als die korrekte orthologe Sequenz angesehen. Da der WU-BLAST-Treffer nicht die gesamte Anfragesequenz abdecken muss, wurden Start und Ende der Anfragesequenz extra-

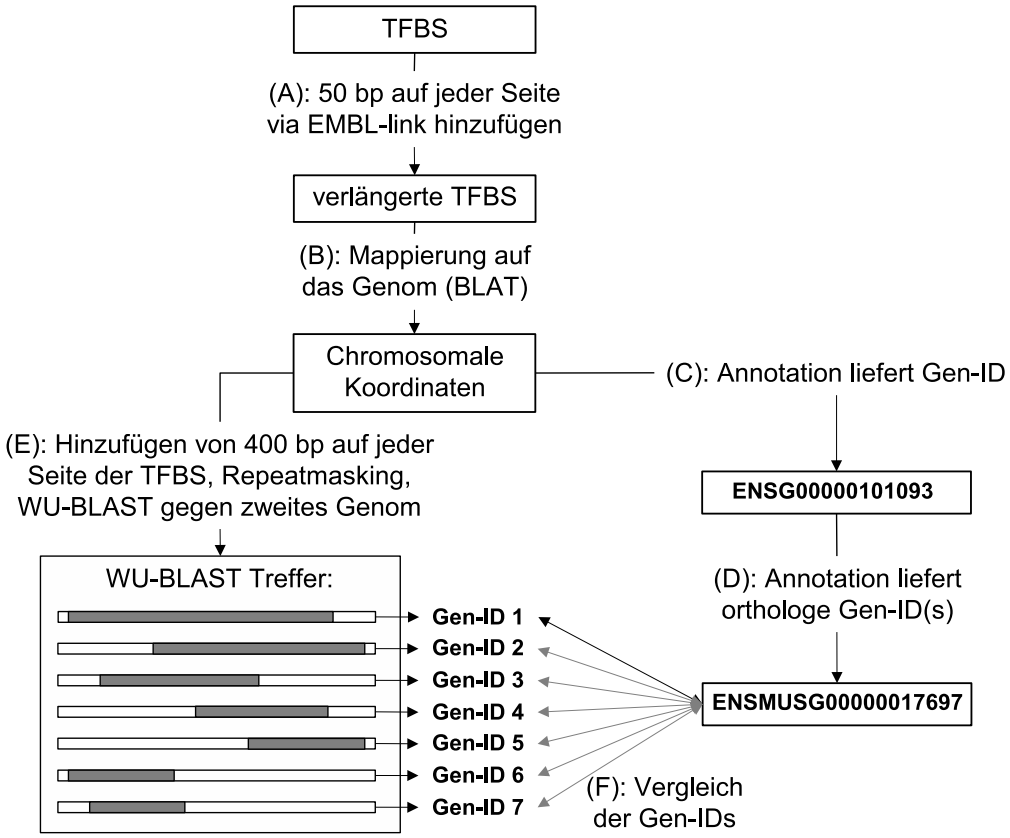

Abbildung 3.1: Die Beschaffung der orthologen Sequenzpaare beginnt mit der Verlängerung der TFBS-Sequenz um 50 bp auf beiden Seiten mittels ihrer EMBL-Links (A) und der anschließenden Mappierung auf das Genom (B). Die TFBS wird dem Gen mit dem nächstgelegenen Transkript zugeordnet (C). Falls orthologe Gene bekannt sind (D), wird die ursprüngliche TFBS-Sequenz um 400 bp auf jeder Seite verlängert, Repeats in ihr maskiert und eine WU-BLAST-Suche gegen das zweite Genom durchgeführt (E). Jeder WU-BLAST-Treffer wird einem Gen zugeordnet (F). Sollte dieses Gen mit einem der bekannten orthologen Gene übereinstimmen, wird dieser WU-BLAST-Treffer als Anker auf dem zweiten Genom akzeptiert.

poliert, um eine orthologe Sequenz ähnlicher Länge zu erhalten (ca. 800 bp). Überlappende orthologe Sequenzpaare (OSPs) wurden in einem OSP zusammengefasst.

3.3.4 Alignment orthologer Sequenzen

Die erhaltenen OSPs wurden mit mehreren globalen Alignment-Algorithmen (AVID, CLUSTAL W, CONREAL, DIALIGN, LAGAN und T-COFFEE) und zwei lokalen Alignment-Algorithmen (BLASTZ und LALIGN) aligniert (siehe 3.2.3, S. 30). Alle Alignment-Programme wurden mit ihren Standard-Parametern benutzt. Bis auf CONREAL ist keines dieser Programme speziell dafür entwickelt worden, nicht-codierende Sequenzen zu ali-

gnieren.

Alle Sequenzen aus Datensatz I, die TFBSs enthalten, wurden auch auf mit BLASTZ angefertigte „whole genome alignments“ (WGAs) mappiert. Diese WGAs wurden von der UCSC-Webseite (ftp://hgdownload.cse.ucsc.edu) heruntergeladen. Die jeweils verwendeten Versionen der WGAs waren „hg17vsMm5“ (Mensch-Maus), „mm5vsHg17“ (Maus-Mensch) und „rn3vsHg17“ (Ratte-Mensch). Die WGAs lagen im sogenannten „axt-Net“-Format vor (http://genome.ucsc.edu/goldenPath/help/axt.html).

3.3.5 Konserviertheitsrate und Hintergrund-Konserviertheit

Die Konserviertheit der TFBSs wurde anhand ihrer Sequenzidentität in den erhaltenen paarweisen Alignments der OSPs berechnet. Die prozentuale Identität (PI) der TFBS i, PI_i (mit $i = 1, \ldots, N$, wobei N der Anzahl der TFBSs entspricht), ist als der Anteil identischer Residuen in den l_i Spalten des paarweisen Alignments, die der der TFBS i entsprechen, definiert (siehe auch Abbildung 3.2):

$$\text{PI}_i = \frac{1}{l_i} \sum_{j=1}^{l_i} a_j, \qquad a_j = \begin{cases} 1, \text{ falls } X_j = Y_j \\ 0, \text{ falls } X_j \neq Y_j \end{cases} \tag{3.5}$$

wobei $X_j, Y_j \in \{\text{A,C,G,T,-}\}$ die alignierten Nukleotide oder Lücken an Position j der TFBS i im paarweisen Alignment sind. Ein Nukleotid, das zu einer Lücke aligniert wurde, wird wie ein Mismatch behandelt. Eine TFBS i wird als konserviert angesehen, wenn ihr PI-Wert größer oder gleich einem bestimmten Konserviertheit-Schwellenwert (KS) ist.

Spalte j	1	2	3	4	5	6	7	8	9	10	11	12
Mensch	G	A	A	A	A	A	C	T	G	T	T	T
Maus	G	-	A	A	A	T	T	T	G	T	T	T
a_j	1	0	1	1	1	0	0	1	1	1	1	1

$$\Rightarrow \text{PI} = \frac{9}{12} = 75\%$$

Abbildung 3.2: Beispiel für die Berechnung des PI-Wertes einer TFBS der Länge $l = 12$ anhand eines paarweisen Alignments.

Die Konserviertheitsrate C_{seq} ist als der Anteil aller N TFBSs, die als konserviert an-

gesehen werden, definiert:

$$C_{seq} = \frac{1}{N}\sum_{i=1}^{N} b_i, \qquad b_i = \begin{cases} 1, \text{ falls PI}_i \geq \text{KS} \\ 0, \text{ falls PI}_i < \text{KS} \end{cases} \tag{3.6}$$

Die Konserviertheit des Hintergrundes wird analog dazu berechnet. Als Hintergrund sind alle Spalten der erhaltenen Alignments definiert, die nicht mit Exons oder bekannten TFBSs überlappen. Jede Länge l_i der gegebenen TFBSs (mit $i = 1, \ldots, N$) wird als sogenanntes „sliding window“ verwendet, um PI-Werte für alle möglichen Positionen des Hintergrundes zu berechnen. Dadurch erhält man eine Verteilung von PI-Werten für diese sogenannten „Hintergrund-Sequenzen“, die auf der gleichen Längenverteilung wie die der PI-Werte zur Berechnung von C_{seq} beruht.

Die Hintergrund-Konserviertheitsrate C_{seq}^{bg} ist als der Anteil aller Hintergrund-Sequenzen definiert, die einen PI-Wert größer oder gleich dem KS besitzen.

3.3.6 Alignments mit randomisierten Transkriptionsfaktor-Bindestellen

Um zu überprüfen, inwiefern die Nukleotidzusammensetzung der TFBSs für ihre Konserviertheit ausschlaggebend ist, wurden in allen OSPs aus Datensatz I die Sequenzen der TFBSs randomisiert, d.h. die Reihenfolge der Nukleotide in einer TFBS wurde zufällig verändert, und die OSPs danach aligniert. Dieser Vorgang wurde 20 mal wiederholt. Die Shuffle-Konserviertheitsrate C_{seq}^{shuf} ist als der Anteil aller randomisierten TFBS-Sequenzen definiert, deren PI-Wert größer oder gleich dem KS ist.

3.3.7 Abhängigkeit der Sequenz-Konserviertheit von der Genfunktion

Um zu überprüfen, ob es einen Zusammenhang zwischen der Funktion eines Gens und der Konserviertheit seiner TFBSs gibt, wurden die TFBSs anhand ihrer zugehörigen Gene in funktionelle Kategorien eingeteilt. Dazu wurde für jede TFBS aus Datensatz I die „gene ontology“-Annotation (GO) (http://www.geneontology.org/) des der TFBS zugewiesenen Gens durch die korrespondierenden „GO slim“-Begriffe ersetzt (Harris et al., 2004). Die „GO slim“-Begriffe bestehen aus Begriffen, die höheren Ebenen der GO-Hierarchie entstammen. Diese wurden vom „European Bioinformatics Institute“ (EBI) ausgewählt, um die meisten Aspekte der Ontologien bezüglich molekularer Funktion, biologischer Prozesse und zellulärer Komponenten abzudecken. Wenn mehrere GO-Begriffe eines Gens dem gleichen „GO slim“-Begriff zugeordnet sind, wurde der entsprechende Begriff nur einmal

gezählt. Die Konserviertheitsrate C_{seq} wurde für alle TFBSs, die mit einem bestimmten „GO slim“-Begriff verknüpft sind, jeweils getrennt berechnet.

3.3.8 Konserviertheitsrate auf Musterebene

Die Konserviertheit der Sequenz einer TFBS besagt nicht zwangsweise, daß die TFBS auch in der orthologen Sequenz aktiv ist, denn durch eine einzelne kritische Mutation könnte die Funktionalität der orthologen TFBS zerstört worden sein. Wenn die DNA-Bindungsspezifität eines TF gering ist, kann andererseits der TF für den Fall einer schwach konservierten TFBS trotzdem an die orthologe Sequenz binden. Es liegt daher nahe, die Konserviertheit einer TFBS auf Musterebene zu bestimmen. Dies wird im Folgenden beschrieben.

Zur Vorhersage von TFBSs in allen OSPs aus Datensatz I mittels des Programms MATCH™ (siehe 3.2.2, S. 29) wurde die PSSM-Vertebraten-Sammlung aus TRANSFAC® benutzt. Alle PSSMs, die TFs repräsentieren, die an eine experimentell bekannte TFBS binden, wurden eingesetzt, um beide Sequenzen eines OSP zu durchsuchen. Die Schwellenwerte für den MSS wurden dem vordefinierten Profil „minFN“, das zur Minimierung falsch negativer Vorhersagen dient, entnommen. Wenn der Start eines PSSM-Treffers in der ersten Sequenz eines OSP im Alignment um nicht mehr als ± 2 bp relativ zum Start eines PSSM-Treffers derselben Matrix in der zweiten Sequenz des OSP verschoben ist und der MSS des Treffers in der ersten Sequenz sich nicht um mehr 0.15 vom MSS des Treffers in der zweiten Sequenz unterscheidet, wird dieser Treffer als „Muster-konserviert“ bezeichnet. Der Start eines PSSM-Treffers ist dabei strangunabhängig immer kleiner als das Ende des PSSM-Treffers. Die Position eines PSSM-Treffers muss nicht zwangsweise mit einer in TRANSFAC® dokumentierten TFBS übereinstimmen, weil in vielen Fällen eine Teilsequenz dieser dokumentierten TFBS zur PSSM-Konstruktion verwendet wurde. Daher wurde eine Positionstoleranz von ± 10 bp um den Start der annotierten TFBS erlaubt, um die Muster-Konserviertheit dieser TFBS zu überprüfen. Liegt der Start eines Muster-konservierten PSSM-Treffers in diesem Bereich, gilt die TFBS als Muster-konserviert. Für den Fall jedoch, dass die TFBS selbst zur Konstruktion der PSSM benutzt wurde, gilt diese Positionstoleranz nicht, denn in diesem Fall ist genau bekannt, wo der PSSM-Treffer erfolgen sollte. Die Konserviertheitsrate C_{pat} ist als der Anteil aller bekannten TFBSs definiert, die Muster-konserviert sind. Die Hintergrund-Konserviertheitsrate C_{pat}^{bg} für einen bestimmten TF ist als der Anteil aller vorhergesagten TFBSs definiert, die Muster-konserviert sind.

3.3.9 Statistische Signifikanz von Konserviertheitsraten

Wenn K TFBSs aus einer Gruppe von N TFBSs konserviert sind, d.h. die Konserviertheitsrate C ist gleich K/N, dann ist die Wahrscheinlichkeit p, diese Konserviertheitsrate C bei einer gegebenen Hintergrund-Konserviertheitsrate C^{bg} zufällig zu beobachten, durch eine kumulative Binomialverteilung gegeben:

$$p = \sum_{j=K}^{N} \binom{N}{j} \left(C^{bg}\right)^{j} \cdot \left(1 - C^{bg}\right)^{(N-j)} \tag{3.7}$$

3.3.10 Prädiktiver Ansatz

Bei der Bestimmung der PI-Werte und damit auch der Konserviertheitsrate C_{seq} der TFBSs wird von den TFBSs selbst ausgegangen (siehe 3.3.5, S. 43), d.h. die Fenstergröße zur Bestimmung des PI-Wertes (Gleichung 3.5) einer TFBS i ist durch deren Länge l_i im Alignment festgelegt.

Ein anderer Ansatz, der sogenannte „prädiktive Ansatz“, ist, das Alignment mit einer fest gewählten Fenstergröße w Position für Position zu überstreichen und für jedes Fenster den zugehörigen PI-Wert zu berechnen. Ein Fenster, das einem bestimmten KS genügt, gilt als konserviert. Die Motivation für diesen Ansatz ist eine optimale Fenstergröße in Kombination mit einem KS zu finden, für die möglichst viele TFBSs und möglichst wenige Hintergrund-Sequenzen von konservierten Fenstern abgedeckt sind. Die Hintergrund-Sequenzen werden für alle Längen l_i (mit $i = 1, \ldots, N$) der alignierten TFBSs aus allen Spalten der erhaltenen Alignments, die nicht mit Exons oder bekannten TFBSs überlappen, gebildet.

Ein Nukleotid kann in maximal w Fenstern enthalten sein. Der maximale PI-Wert dieser w Fenster wird diesem Nukleotid zugeordnet. Wenn dieser PI-Wert größer oder gleich dem gewählten KS ist, gilt das Nukleotid als konserviert. Jeder TFBS oder Hintergrund-Sequenz wird nun der minimale PI-Wert von allen Nukleotiden zugewiesen, die sie überdeckt (siehe Abbildung 3.3). Dies bedeutet, dass eine TFBS oder Hintergrund-Sequenz konserviert ist, wenn dieser minimale PI-Wert dem KS genügt. In diesem Fall besteht sie komplett aus konservierten Nukleotiden, oder wird, anders ausgedrückt, komplett von konservierten Fenstern überdeckt. Die Konserviertheitsraten C_{seq} und C_{seq}^{bg} werden analog wie in 3.3.5 berechnet.

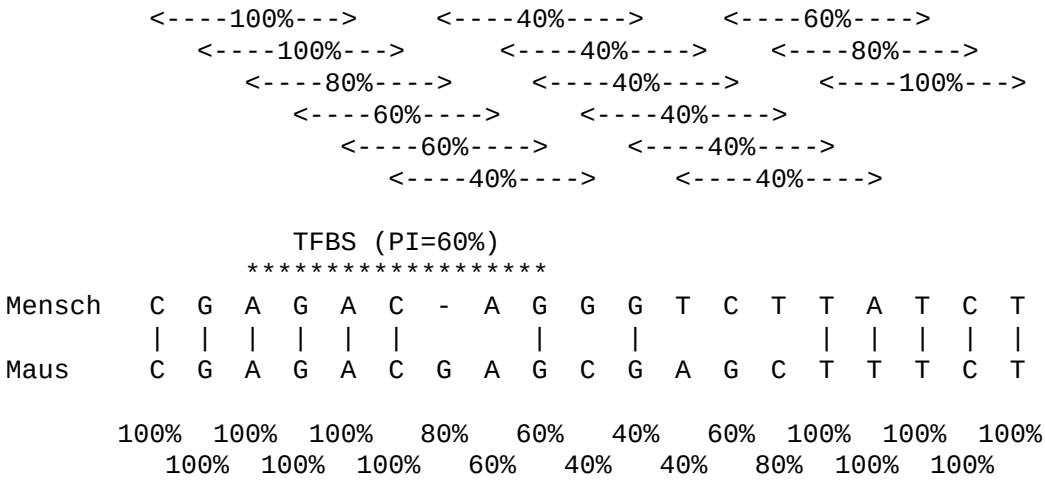

Abbildung 3.3: In diesem Beispiel wird das paarweise Alignment mit einer Fenstergröße $w = 5$ bp Position für Position überstrichen und der PI-Wert jedes Fensters berechnet. Jedem Nukleotid wird der maximale PI-Wert zugeordnet, den eines der Fenster, in denen es enthalten ist, besitzt. Die TFBS in diesem Beispiel besitzt die Länge $l = 7$ bp. Die PI-Werte der Nukleotide, die von der TFBSs überdeckt werden, reichen von 60 bis 100%. Der TFBS wird der minimale PI-Wert von 60% zugeordnet, d.h. wenn man das Alignment mit einer Fenstergröße $w = 5$ bp untersucht, ist die TFBS nur bei einem KS $\leq 60\%$ vollständig von konservierten Fenstern bedeckt.

3.4 Ein Hidden-Markov-Modell zur Vorhersage von Transkriptionsfaktor-Bindestellen

PSSMs und phylogenetisches Footprinting sind zwei voneinander unabhängige Methoden, um TFBSs in nicht-codierenden Regionen vorherzusagen. Eine Kombination dieser Methoden sollte Vorteile bei der Vorhersage von potentiellen TFBSs bringen. Die einfachste Kombination ist, die Suche mit PSSMs auf konservierte Bereiche, die mittels des prädiktiven Ansatzes bestimmt wurden (siehe 3.3.10, S. 46), einzuschränken, um falsch positive Vorhersagen zu minimieren. Allerdings stellt sich bei diesem Ansatz die Frage, wie man die Schwellenwerte für die Konserviertheit, den MSS oder für eine lineare Kombination aus beiden (siehe 3.4.7, S. 62) wählt.

Ein anderer Ansatz ist die Kombination beider Methoden in einem HMM, das auf einer Eingabesequenz TFBSs eines bestimmten TF möglichst zuverlässig vorhersagen soll. Hier „entscheidet“ das HMM, welchen Beitrag das Profil der PSSM und die Konserviertheit der entsprechenden Sequenz zur Vorhersage einer TFBS leisten. Die Wahl von zwei Schwellenwerten entfällt hierdurch. Dies ermöglicht einerseits, dass eine TFBS erkannt wird, wenn sie nicht konserviert ist, aber dem Profil der PSSM sehr gut entspricht. Andererseits kann eine TFBS auch dann erkannt werden, wenn sie dem Profil nur moderat ähnelt, aber hoch konserviert ist. Solche TFBSs würden nicht erkannt werden, wenn man nur in hoch konservierten Bereichen nach PSSM-Treffern mit hohem MSS sucht.

3.4.1 Formale Beschreibung eines Hidden-Markov-Modells

Im Folgenden wird ein HMM formal beschrieben. Ein HMM ist ein probabilistisches Modell, das durch zwei gekoppelte Zufallsprozesse beschrieben werden kann. Der erste Zufallsprozess ist eine Markov-Kette, die durch Zustände und Übergänge zwischen diesen Zuständen gekennzeichnet ist.

3.4.1.1 Definition einer Markov-Kette

Eine Markov-Kette besteht aus mehreren Zuständen, die durch Übergänge miteinander verbunden sind. Jeder Übergang erfolgt mit einer gewissen Wahrscheinlichkeit, der sogenannten Übergangswahrscheinlichkeit. Für eine „homogene“ Markov-Kette sind die Übergangswahrscheinlichkeiten unabhängig vom Zeitpunkt.

Im Folgenden betrachten wir Markov-Ketten 1. Ordnung, d.h. die Wahrscheinlichkeit eines jeden Zustandes x_i, den die Markov-Kette zum Zeitpunkt i annimmt, hängt nur vom

Zustand x_{i-1} zum Zeitpunkt $i-1$ ab, jedoch nicht von der gesamten vorherigen Zustandsfolge. D.h. es gilt:

$$P(x_i|x_{i-1},\ldots,x_1) = P(x_i|x_{i-1}) = a_{x_{i-1}x_i} \tag{3.8}$$

Die Wahrscheinlichkeit einer Folge $x = x_1, x_2, \ldots, x_L$ von Zuständen der Länge L ist daher gegeben durch:

$$\begin{aligned} P(x) &= P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})\cdots P(x_2|x_1)P(x_1) \\ &= P(x_1)\prod_{i=2}^{L} a_{x_{i-1}x_i} \end{aligned} \tag{3.9}$$

Um zu modellieren, mit welcher Wahrscheinlichkeit $P(x_1)$ eine Markov-Kette in einem bestimmten Zustand x_1 beginnt, wird dem Modell ein Anfangszustand $x_0 = \mathcal{B}$ hinzugefügt. Die Wahrscheinlichkeit, mit der zum Zeitpunkt 1 der Zustand s angenommen wird, ist gegeben durch:

$$P(x_1 = s) = a_{\mathcal{B}s} \tag{3.10}$$

Analog kann dem Modell ein Endzustand $x_{L+1} = \mathcal{E}$ hinzugefügt werden. Die Übergangswahrscheinlichkeit vom Zustand t in den Endzustand $\mathcal{E}$ ist gegeben durch:

$$P(\mathcal{E}|x_L = t) = a_{t\mathcal{E}} \tag{3.11}$$

3.4.1.2 Definition eines Hidden-Markov-Modells

In einem HMM ist jeder Zustand einer Markov-Kette mit einem zweiten Zufallsprozess verknüpft, der nach einer dem Zustand zugehörigen Emissions-Wahrscheinlichkeitsverteilung eine Beobachtung emittiert. Für ein HMM gibt es keine eindeutige Zuordnung zwischen den beobachteten Symbolen und den zugrunde liegenden Zuständen, denn es muss zwischen der Folge der Zustände und der Folge der Symbole unterschieden werden. Die Folge der Zustände wird Zustandspfad π genannt, wobei der i-te Zustand des Zustandspfads mit π_i bezeichnet wird. Die Wahrscheinlichkeit, zum Zeitpunkt i in einem Zustand l zu sein, hängt für eine Markov-Kette 1. Ordnung nur vom Zustand k ab, der zum Zeitpunkt $i-1$ angenommen wird:

$$a_{kl} = P(\pi_i = l|\pi_{i-1} = k) \tag{3.12}$$

Für die Zustandsfolge des HMM gibt es einen Anfangs- und Endzustand. Z.B. gibt die Übergangswahrscheinlichkeit $a_{\mathcal{B}k}$ die Wahrscheinlichkeit an, dass zum Zeitpunkt 1 der Zustand k angenommen wird.

Jeder Zustand k emittiert ein Symbol b aus einem Emissionsalphabet Σ entsprechend einer gegebenen Emissionswahrscheinlichkeit $e_k(b)$. Diese gibt an, mit welcher Wahrscheinlichkeit man in einem Zustand k ein Symbol b beobachtet:

$$e_k(b) = P(s_i = b|\pi_i = k), \qquad b \in \Sigma \tag{3.13}$$

Ein HMM kann benutzt werden, um eine Folge von Symbolen zu erzeugen: Der erste Zustand π_1 wird entsprechend der Übergangswahrscheinlichkeit $a_{\mathcal{B}\pi_1}$ angenommen. Aus diesem Zustand wird ein Symbol entsprechend der Emissionswahrscheinlichkeitsverteilung e_{π_1} dieses Zustandes emittiert. Im nächsten Schritt wird ein Zustand π_2 entsprechend der Übergangswahrscheinlichkeiten $a_{\pi_1\pi_2}$ angenommen, der wiederum ein Symbol entsprechend e_{π_2} emittiert. Auf diese Weise kann eine artifizielle Sequenz erzeugt werden. Die gemeinsame Wahrscheinlichkeit einer Sequenz s und eines Zustandspfads π ist daher gegeben durch:

$$P(s, \pi) = a_{\mathcal{B}\pi_1} \prod_{i=1}^{L} e_{\pi_i}(s_i) a_{\pi_i\pi_{i+1}} \tag{3.14}$$

wobei $\pi_{L+1} = \mathcal{E}$ gilt.

3.4.1.3 Viterbi-Algorithmus

Hat man eine Sequenz s gegeben, stellt sich die Frage, welcher Zustandspfad π dieser Sequenz zugrunde liegt. Es gibt mehrere Ansätze, um den Zustandspfad abzuschätzen, wobei der bekannteste eine Methode des „Dynamischen Programmierens“ ist, nämlich der sogenannte „Viterbi-Algorithmus“ (Viterbi, 1967).

Eine beobachtete Sequenz s kann durch viele Zustandspfade π erzeugt werden, wobei sich aber die zugehörigen Wahrscheinlichkeiten, die nach Gleichung (3.14) berechnet werden, erheblich unterscheiden können. Wenn man sich für einen Pfad entscheiden muss, der die Beobachtung erklärt, ist es naheliegend, den Pfad mit der maximalen Wahrscheinlichkeit zu wählen:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, P(s, \pi) \tag{3.15}$$

Der wahrscheinlichste Pfad π^*, der sogenannte „Viterbi-Pfad“, kann durch Rekursion bestimmt werden: Wenn die Wahrscheinlichkeit $v_k(i)$, die die „Viterbi-Variable“ des

Zustandes k zum Zeitpunkt i genannt wird, für den wahrscheinlichsten Pfad, der zum Zeitpunkt i im Zustand k endet und dabei das Präfix $s_1, \ldots, s_i$ der Sequenz s emittiert, für alle Zustände k bekannt ist, kann die Viterbi-Variable $v_l(i+1)$ für den wahrscheinlichsten Pfad, der zum Zeitpunkt $i+1$ im Zustand l endet und dabei das Präfix $s_1, \ldots, s_{i+1}$ der Sequenz s emittiert, folgendermaßen berechnet werden:

$$v_l(i+1) = e_l(s_{i+1}) \max_k \left(v_k\left(i\right) a_{kl}\right) \tag{3.16}$$

Da alle Zustandspfade im Anfangszustand $\mathcal{B}$ beginnen, gilt zum Zeitpunkt $i = 0$ die Anfangsbedingung $v_{\mathcal{B}}(0) = 1$. Die Viterbi-Variable für einen bestimmten Zustand l zum Zeitpunkt $i+1$ wird nach Gleichung (3.16) also aus zwei Faktoren berechnet. Der erste Faktor ist die Emissionswahrscheinlichkeit $e_l(s_{i+1})$, die Beobachtung s_{i+1} aus dem Zustand l zu emittieren. Der zweite Faktor ist das maximale Produkt aus der Viterbi-Variable $v_k(i)$, d.h. der Wahrscheinlichkeit des wahrscheinlichsten Pfades, der mit Beobachtung s_i im Zustand k endet, und der Übergangswahrscheinlichkeit a_{kl}, um vom Zustand k in den Zustand l zu gelangen.

Für jeden Zustand l wird ein Zeiger $\text{ptr}_i(l)$ auf denjenigen vorhergehenden Zustand k gerichtet, für den das Produkt $v_k\left(i\right) \cdot a_{kl}$ maximal ist. Der Viterbi-Pfad π^* wird ermittelt, indem die Zeiger ausgehend von dem Zustand z, für den die Viterbi-Variable zum Zeitpunkt L, $v_z(L)$, maximal wird, zurückverfolgt werden („backtracking“). Der vollständige Algorithmus lautet:

Initialisierung ($i = 0$):	$v_{\mathcal{B}}(0) = 1$, $v_k(0) = 0$ für alle Zustände $k \neq \mathcal{B}$.
Rekursion ($i = 1 \ldots L$):	$v_l(i) = e_l(s_i) \max_k \left(v_k(i-1)a_{kl}\right)$; $\text{ptr}_i(l) = \text{argmax}_k \left(v_k(i-1)a_{kl}\right)$.
Termination:	$P(s, \pi^*) = \max_k(v_k(L)a_{k\mathcal{E}})$; $\pi^*_L = \text{argmax}_k(v_k(L)a_{k\mathcal{E}}$.
Zurückverfolgung ($i = L \ldots 1$):	$\pi^*_{i-1} = \text{ptr}_i(\pi^*_i)$

3.4.2 Zustände und Übergänge des Hidden-Markov-Modells

Im Rahmen dieser Arbeit wurde ein HMM entwickelt, welches funktionelle Bereiche, die TFBSs eines bestimmten TF repräsentieren, von solchen unterscheiden soll, die nicht-funktionell oder aber konserviert, jedoch nicht repräsentativ für den untersuchten TF sind.

Als Informationsquellen nutzt das HMM die PSSM eines bestimmten TF und ein Alignment orthologer nicht-codierender Sequenzen. Das HMM besteht aus mehreren Zuständen, die die oben genannten Anforderungen abdecken sollen:

1. Der Zustand **NB** emittiert Nukleotide und Lücken in nicht-funktionellen Bereichen.
2. Die Zustände $\mathbf{B_1^+}, \mathbf{B_2^+}, \ldots, \mathbf{B_\lambda^+}$ emittieren die einzelnen Positionen einer TFBS der Länge λ auf dem Vorwärts-Strang.
3. Die Zustände $\mathbf{G_1^+}, \mathbf{G_2^+}, \ldots, \mathbf{G_{\lambda\text{-}1}^+}$ emittieren Lücken zwischen den einzelnen Positionen einer TFBS der Länge λ auf dem Vorwärts-Strang.
4. Die Zustände $\mathbf{B_1^-}, \mathbf{B_2^-}, \ldots, \mathbf{B_\lambda^-}$ emittieren die einzelnen Positionen einer TFBS der Länge λ auf dem Rückwärts-Strang.
5. Die Zustände $\mathbf{G_1^-}, \mathbf{G_2^-}, \ldots, \mathbf{G_{\lambda\text{-}1}^-}$ emittieren Lücken zwischen den einzelnen Positionen einer TFBS der Länge λ auf dem Rückwärts-Strang.
6. Der Zustand **F** emittiert Nukleotide und Lücken in konservierten Bereichen, die nicht dem Muster des TF entsprechen.
7. Der sogenannte „Magical State“ (**MS**) vereinigt den obligatorischen Anfangs- und Endzustand des HMM in einem Zustand und ist eine Besonderheit der Implementation von HMMs in BioJava.

Abbildung 3.4 zeigt eine Übersicht über die möglichen Übergänge zwischen den einzelnen Zuständen. Die dazugehörigen Übergangswahrscheinlichkeiten lassen sich Tabelle 3.4 entnehmen.

Tabelle 3.4: Übergangswahrscheinlichkeiten des HMM

Zustand	**MS**	**NB**	**F**	$\mathbf{B_1^+}$	$\mathbf{B_1^-}$	$\mathbf{G_1^+}$	$\mathbf{G_1^-}$	$\mathbf{B_2^+}$	$\mathbf{B_2^-}$	$\cdots$
MS	0	$(1-f)(1-p)$	$f(1-p)$	$p/2$	$p/2$	0	0	0	0	$\cdots$
NB	$p/3$	$(1-f)(1-p)$	$f(1-p)$	$p/3$	$p/3$	0	0	0	0	$\cdots$
F	$p/3$	$(1-f)(1-p)$	$f(1-p)$	$p/3$	$p/3$	0	0	0	0	$\cdots$
$\mathbf{B_1^+}$	0	0	0	0	0	g	0	$1-g$	0	$\cdots$
$\mathbf{B_1^-}$	0	0	0	0	0	0	g	0	$1-g$	$\cdots$
$\mathbf{G_1^+}$	0	0	0	0	0	g	0	$1-g$	0	$\cdots$
$\mathbf{G_1^-}$	0	0	0	0	0	0	g	0	$1-g$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
$\mathbf{B_\lambda^+}$	$p/3$	$(1-f)(1-p)$	$f(1-p)$	$p/3$	$p/3$	0	0	0	0	$\cdots$
$\mathbf{B_\lambda^-}$	$p/3$	$(1-f)(1-p)$	$f(1-p)$	$p/3$	$p/3$	0	0	0	0	$\cdots$

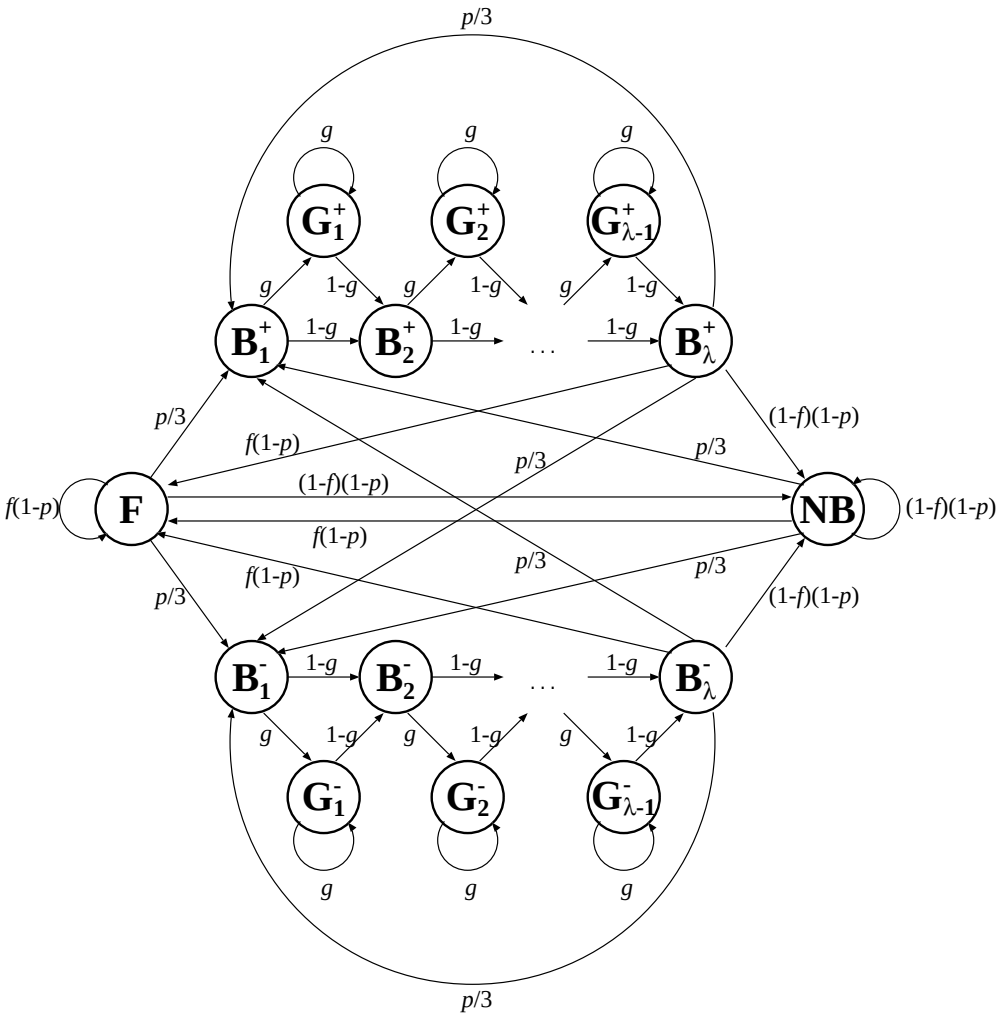

Abbildung 3.4: Übersicht über die Zustände des HMMs und die Übergänge, dargestellt durch Pfeile, zwischen diesen. Der Übersichtlichkeit halber ist der sogenannte „Magical State“ (**MS**) nicht eingezeichnet. Die Beschriftungen an den Pfeilen entsprechen den Übergangswahrscheinlichkeiten.

Der Parameter $p \in [0,1]$ in Tabelle 3.4 beeinflusst die Sensitivität bei der Vorhersage von TFBSs. Je höher die Wahrscheinlichkeit p ist, um so häufiger findet ein Übergang in den ersten Zustand einer TFBS statt. Dieser Übergang ist aus den Zuständen **NB**, **F**, $\mathbf{B}_\lambda^+$ oder $\mathbf{B}_\lambda^-$ möglich und findet mit einer Wahrscheinlichkeit von $p/3$ statt. Weiterhin ist ein Übergang in den Zustand **MS** mit einer Wahrscheinlichkeit von $p/3$ möglich. Aus den Zuständen $\mathbf{B}_\mathbf{1}^+$ bzw. $\mathbf{B}_\mathbf{1}^-$ wird jeweils das erste Nukleotid einer TFBS auf dem Vorwärts- bzw. Rückwärts-Strang emittiert. Ausgehend von $\mathbf{B}_\mathbf{1}^+$ bzw. $\mathbf{B}_\mathbf{1}^-$ werden anschließend Zustände durchlaufen, die die restlichen Nukleotide der TFBS bzw. Lücken emittieren. Für jeden Zustand $\mathbf{B}_\mathbf{i}^+$ bzw. $\mathbf{B}_\mathbf{i}^-$ ($i = 1, \ldots, \lambda - 1$) des HMM existiert ein Übergang in den Zustand $\mathbf{B}_\mathbf{i+1}^+$ bzw. $\mathbf{B}_\mathbf{i+1}^-$, der das Nukleotid der nächsten Position $i + 1$ emittiert. Die Wahrschein-

lichkeit für diesen Übergang ist $(1-g)$, wobei $g \in [0,1]$ die Wahrscheinlichkeit ist, nach Position i der TFBS eine Lücke einzufügen. g ist somit die Übergangswahrscheinlichkeit für einen Übergang vom Zustand $\mathbf{B_i^+}$ bzw. $\mathbf{B_i^-}$ in den Zustand $\mathbf{G_i^+}$ bzw. $\mathbf{G_i^-}$, der eine Lücke emittiert. Dieser Zustand $\mathbf{G_i^+}$ bzw. $\mathbf{G_i^-}$ wird mit Wahrscheinlichkeit g beibehalten, oder mit Wahrscheinlichkeit $(1-g)$ der folgende Zustand $\mathbf{B_{i+1}^+}$ bzw. $\mathbf{B_{i+1}^-}$, der das nächste Nukleotid der TFBS emittiert, angenommen.

Die Zustände **NB**, **F**, $\mathbf{B_\lambda^+}$ und $\mathbf{B_\lambda^-}$ besitzen also jeweils drei Übergänge mit einer Wahrscheinlichkeit von $p/3$, den jeweiligen Zustand zu verlassen. Es verbleibt noch die Übergangswahrscheinlichkeit $(1-p)$, vom jeweiligen Zustand in die Zustände **F** oder **NB** überzugehen bzw. in diesen zu verbleiben. Diese Wahrscheinlichkeit wird mittels des Parameters $f \in [0,1]$ auf die Zustände **F** und **NB** aufgeteilt, d.h. die Wahrscheinlichkeit, in den Zustand **F** zu gelangen, beträgt $f(1-p)$, und die Wahrscheinlichkeit, in den Zustand **NB** zu gelangen, ist $(1-f)(1-p)$.

3.4.3 Emissionen des Hidden-Markov-Modells

Aus den Zuständen des HMM können konservierte Nukleotide (symbolisiert durch $\Sigma^k = \{A, C, G, T\}$), nicht konservierte Nukleotide ($\Sigma^n = \{a, c, g, t\}$) und Lücken ($-$) emittiert werden. Das Emissionsalphabet $\Sigma = \Sigma^k \cup \Sigma^n \cup \{-\}$ besteht daher aus neun Symbolen.

Die einzelnen Zustände k des HMM emittieren mit unterschiedlichen Wahrscheinlichkeiten $e_k(b)$ Symbole b aus dem Emissionsalphabets Σ. Tabelle 3.5 gibt eine Übersicht über die Parameter, die das Emissionsverhalten der einzelnen Zustände bestimmen. Diese Parameter wurden empirisch ermittelt.

Tabelle 3.5: Parameter zur Berechnung der Emissionswahrscheinlichkeiten des HMM

Parameter	Beschreibung
c_{tfbs}	Wahrscheinlichkeit, in den Zuständen $\mathbf{B_1^+}$, $\mathbf{B_2^+}$,..., $\mathbf{B_\lambda^+}$ bzw. $\mathbf{B_1^-}$, $\mathbf{B_2^-}$,..., $\mathbf{B_\lambda^-}$ ein konserviertes Nukleotid zu emittieren
$h_{b,i}$	Wahrscheinlichkeit, in einem Zustand $\mathbf{B_i^+}$ bzw. $\mathbf{B_i^-}$ (mit $i = 1, \ldots, \lambda$) ein bestimmtes Nukleotid b zu emittieren
c_{nb}	Wahrscheinlichkeit, dass ein im Zustand **NB** emittiertes Nukleotid konserviert ist
g_{nb}	Wahrscheinlichkeit, eine Lücke aus dem Zustand **NB** zu emittieren
c_{f}	Wahrscheinlichkeit, dass ein im Zustand **F** emittiertes Nukleotid konserviert ist
g_{f}	Wahrscheinlichkeit, eine Lücke aus dem Zustand **F** zu emittieren

3.4.4 Bestimmung der Parameter des Hidden-Markov-Modells

Die Parameter c_{tfbs} und g wurden für jede untersuchte PSSM empirisch aus den vorliegenden Mensch-Maus-Alignments aus Datensatz I ermittelt. Für eine PSSM der Länge λ, die durch N TFBS im Datensatz vertreten ist, erhält man:

$$g = \frac{a}{1+a}, \text{ mit} \qquad a = \frac{1}{N \cdot (\lambda - 1)} \sum_{j=1}^{N} L_j \tag{3.17}$$

$$c_{\text{tfbs}} = \frac{1}{N \cdot \lambda} \sum_{j=1}^{N} M_j \tag{3.18}$$

Dabei ist L_j die Anzahl der Lücken und M_j die Anzahl der konservierten Nukleotide in der j-ten TFBS.

Die korrigierte relative Häufigkeit $h_{b,i}$ eines bestimmten Nukleotids $b \in \{\text{A,C,G,T}\}$ an Position i der PSSM, wird berechnet durch:

$$h_{b,i} = \frac{n_{b,i} + \alpha_b}{\sum_{a \in \{\text{A,C,G,T}\}} (n_{a,i} + \alpha_a)} \tag{3.19}$$

Dabei ist $n_{b,i}$ die Häufigkeit von Nukleotid b an Position i der eingesetzten PSSM und α_b der Pseudocount für Nukleotid b. Es wurde ein Wert von $\alpha_b = 2$ verwendet.

Damit ergibt sich z.B. für die Emissionswahrscheinlichkeit $e_{\mathbf{B_i^+}}(b)$ eines konservierten Nukleotids $b \in \Sigma^k$ im Zustand $\mathbf{B_i^+}$ (mit $i = 1, \ldots, \lambda$):

$$e_{\mathbf{B_i^+}}(b) = h_{b,i} \cdot c_{\text{tfbs}}, \qquad b \in \Sigma^k \tag{3.20}$$

und analog für ein nicht konserviertes Nukleotid $b \in \Sigma^n$:

$$e_{\mathbf{B_i^+}}(b) = h_{b,i} \cdot (1 - c_{\text{tfbs}}), \qquad b \in \Sigma^n \tag{3.21}$$

In den Zuständen **NB** und **F** besitzen alle Nukleotide $b \in \{\text{A,C,G,T}\}$ die gleiche relative Häufigkeit von $h_b = 0.25$. Die Wahrscheinlichkeit c_{f}, dass ein im Zustand **F** emittiertes Nukleotid konserviert ist, wurde analog zu Gleichung (3.18) aus dem relativen Anteil konservierter Nukleotide in allen untersuchten TFBSs berechnet:

$$c_{\text{f}} = \sum_{j=1}^{N} \frac{M_j}{\lambda_j} \tag{3.22}$$

λ_j ist die Länge der j-ten TFBS und M_j die Anzahl der konservierten Nukleotide in der j-ten TFBS. Die Wahrscheinlichkeit g_{f} im Zustand **F** eine Lücke zu emittieren, berechnet sich analog dazu aus dem relativen Anteil aller Lücken in allen N TFBSs:

$$g_{\text{f}} = \sum_{j=1}^{N} \frac{L_j}{L_j + \lambda_j} \tag{3.23}$$

Dabei ist L_j die Anzahl der Lücken der j-ten TFBS. Der relative Anteil der konservierten Nukleotide an allen Nukleotiden des Hintergrundes, d.h. aller Nukleotide, die nicht zu einer bekannten TFBS oder einem Exon gehören, ist die Wahrscheinlichkeit c_{nb}, dass ein im Zustand **NB** emittiertes Nukleotid konserviert ist. Analog erhält man die Wahrscheinlichkeit g_{nb}, im Zustand **NB** eine Lücke zu emittieren, aus dem relativen Anteil der Lücken an allen Alignment-Positionen des Hintergrundes.

Tabelle 3.6 zeigt eine Übersicht der Emissionswahrscheinlichkeiten der einzelnen Zustände.

Tabelle 3.6: Emissionswahrscheinlichkeiten des HMM

	NB	**F**	$\mathbf{B_i^+}$	$\mathbf{B_i^-}$	$\mathbf{G_i}$
A	$(1-g_{\text{nb}})c_{\text{nb}}/4$	$(1-g_{\text{f}})c_{\text{f}}/4$	$h_{A,i}c_{\text{tfbs}}$	$h_{T,\lambda+1-i}c_{\text{tfbs}}$	0
C	$(1-g_{\text{nb}})c_{\text{nb}}/4$	$(1-g_{\text{f}})c_{\text{f}}/4$	$h_{C,i}c_{\text{tfbs}}$	$h_{G,\lambda+1-i}c_{\text{tfbs}}$	0
G	$(1-g_{\text{nb}})c_{\text{nb}}/4$	$(1-g_{\text{f}})c_{\text{f}}/4$	$h_{G,i}c_{\text{tfbs}}$	$h_{C,\lambda+1-i}c_{\text{tfbs}}$	0
T	$(1-g_{\text{nb}})c_{\text{nb}}/4$	$(1-g_{\text{f}})c_{\text{f}}/4$	$h_{T,i}c_{\text{tfbs}}$	$h_{A,\lambda+1-i}c_{\text{tfbs}}$	0
a	$(1-g_{\text{nb}})(1-c_{\text{nb}})/4$	$(1-g_{\text{f}})(1-c_{\text{f}})/4$	$h_{A,i}(1-c_{\text{tfbs}})$	$h_{T,\lambda+1-i}(1-c_{\text{tfbs}})$	0
c	$(1-g_{\text{nb}})(1-c_{\text{nb}})/4$	$(1-g_{\text{f}})(1-c_{\text{f}})/4$	$h_{C,i}(1-c_{\text{tfbs}})$	$h_{G,\lambda+1-i}(1-c_{\text{tfbs}})$	0
g	$(1-g_{\text{nb}})(1-c_{\text{nb}})/4$	$(1-g_{\text{f}})(1-c_{\text{f}})/4$	$h_{G,i}(1-c_{\text{tfbs}})$	$h_{C,\lambda+1-i}(1-c_{\text{tfbs}})$	0
t	$(1-g_{\text{nb}})(1-c_{\text{nb}})/4$	$(1-g_{\text{f}})(1-c_{\text{f}})/4$	$h_{T,i}(1-c_{\text{tfbs}})$	$h_{A,\lambda+1-i}(1-c_{\text{tfbs}})$	0
-	g_{nb}	g_{f}	0	0	1

3.4.5 Implementation des Hidden-Markov-Modells

Die Implementation des HMM erfolgte mittels des Paketes „org.biojava.bio.dp“ aus BioJava 1.4 (http://www.biojava.org), das Klassen für HMMs und Algorithmen des dynamischen Programmierens zur Verfügung stellt.

Eingabe

Das HMM benötigt als Eingabe u.a. eine Datei, die ein paarweises Alignment im Multi-FASTA-Format enthält. Jeder der beiden Einträge dieser Datei besteht aus einer beschrei-

benden Kopfzeile beginnend mit dem Größer-als-Zeichen ('>'), und der eigentlichen Sequenz in den darauffolgenden Zeilen. Abbildung 3.5 zeigt ein Beispiel einer solchen Datei.

```
>human
AGACACGAAGAGTCTGAGCATCTATAAACAGCAACGGAAGAAATGAAATTG
GCTGCGTCCTCTAAGCCTGTCCCCGCAGCATGCTGGAGGAGGGTCGCGGGG
GACATGGAAGAGGAGGAGCTTTGGAGAGAGGATGCTTGTGCTCCCCCGCCT
TTTCTTGCTATTTCTATTTGGGGGTTGGATTCCTGGGAGCTTCATCACATT
>mouse
ACCACCACATATTTCATGCTCAGAGTCTGGGCTTTGGAAAGAAGAAAAGTG
GTTTCATCCTCCAAACCCATCCTGTTGGCACCTTGGACACGAGTCTTGGGA
GGCAGTAAAAATAAA---CTTTAGAGGAAA------TATGCTCTTCTAACA
TCCCTTGGTGTTTACACGTGTAGCTGAATTCTTTAGAGCCTGTATCATGTT
```

Abbildung 3.5: Beispiel eines paarweisen Alignments im Multi-FASTA-Format. Jeder alignierten Sequenz geht eine beschreibende Zeile voraus, die als erstes Symbol ein Größer-als-Zeichen ('>') enthält.

Das paarweise Alignment wird nach dem Einlesen auf eine einzelne Zeichenkette s, basierend auf dem Emissionsalphabet Σ, reduziert. Dazu wird die Maus-Sequenz auf die Mensch-Sequenz „projiziert", d.h. die Maus-Sequenz liefert nur die Information, ob Nukleotide der Mensch-Sequenz konserviert oder nicht konserviert sind: Nukleotide der Mensch-Sequenz, die zu identischen Nukleotiden in der Maus-Sequenz aligniert sind, werden als konservierte Nukleotide („Matches") repräsentiert. Die übrigen Nukleotide der Mensch-Sequenz werden als nicht konservierte Nukleotide („Mismatches") behandelt. Lücken in der Mensch-Sequenz bleiben unverändert (siehe Abbildung 3.6).

```
Mensch   T A A G A C T T A - - T A T T A
Maus     T A G T C C C T A G C T A T T -

=> Ergebnis T A a g a C t T A - - T A T T a
```

Abbildung 3.6: Reduktion eines paarweisen Alignments auf eine Zeichenkette s, die aus konservierten Nukleotiden (A, C, G, T), nicht konservierten Nukleotiden (a, c, g, t) und Lücken ($-$) besteht. Die Mensch-Sequenz wird als Bezugspunkt verwendet, d.h. die Maus-Sequenz liefert nur die Information, ob Nukleotide der Mensch-Sequenz konserviert oder nicht konserviert sind.

Als zweite Eingabe ist eine Datei erforderlich, die die Häufigkeitsverteilungen der Nukleotide innerhalb einer PSSM enthält. Sie besteht aus einer ersten Zeile, die die TRANSFAC®-Zugriffsnummer der PSSM enthält, einer Trennzeile ('//') und darauffolgend je einer Zeile pro Position der PSSM. Jede dieser zuletzt genannten Zeilen enhält vier durch

Tabulatoren getrennte Einträge, die die Häufigkeit der Nukleotide A, C, G und T repräsentieren (siehe Abbildung 3.7). Diese Einträge werden zur Berechnung der relativen Häufigkeit $h_{b,i}$ eines bestimmten Nukleotids b an Position i der PSSM mittels Gleichung (3.19) benötigt.

```
AC  M00926
//
8       7       19      85
27      5       65      22
76      13      13      17
6       17      86      10
0       1       0       118
4       112     1       2
118     1       0       0
12      29      30      48
```

Abbildung 3.7: Beispieldatei für die Nukleotidhäufigkeits-Verteilung einer PSSM. Die Datei enthält die TRANSFAC®-Zugriffsnummer der PSSM, gefolgt von einer Trennzeile ('//') und den Häufigkeiten der einzelnen Nukleotide an jeder Position in der Reihenfolge A, C, G, T.

Als weitere Eingabe ist ein Wert für die Übergangswahrscheinlichkeit p zwischen 0 und 1 erforderlich. Zusätzlich können die Parameter aus Tabelle 3.5 dem Programm übergeben werden, wodurch die voreingestellten Standardwerte überschrieben werden.

Ausgabe

Mittels des Viterbi-Algorithmus (siehe 3.4.1.3, S. 50) wird der wahrscheinlichste Zustandspfad π, der der aus dem paarweisen Alignment erhaltenen Sequenz s zugrunde liegt, ermittelt. Alle Abschnitte der Mensch-Sequenz, die der Viterbi-Algorithmus den Zuständen $\mathbf{B_1^+}, \mathbf{B_2^+}, \ldots, \mathbf{B_\lambda^+}$ bzw. $\mathbf{B_1^-}, \mathbf{B_2^-}, \ldots, \mathbf{B_\lambda^-}$ zuordnet, werden vom Programm als vorhergesagte TFBSs ausgegeben.

Die Ausgabe des Programms erfolgt im „General Feature Format“ (GFF) , das Abbildung 3.8 beispielhaft zeigt (http://www.sanger.ac.uk/Software/formats/GFF/).

Kommentarzeilen beginnen mit einem '#'-Zeichen. Alle Parameter, die dem Programm übergeben werden und von den Standardwerten abweichen, werden in Kommentarzeilen ausgegeben. Jede weitere Zeile entspricht einer vorhergesagten TFBS. Dabei gibt die erste Spalte den Namen der untersuchten Sequenz an, die zweite Spalte den Namen des Programms, das die Vorhersage gemacht hat, und die dritte Spalte eine Klassifikation der Vorhersage. In der vierten und fünften Spalte sind der Start- und Endpunkt der Vorhersage auf der Sequenz angegeben. Diese Ausgabe erfolgt bezogen auf die Positionen in der Sequenz,

```
# p = 0.02
human   HMM   tfbs   772    779    .   -   .   M00926
human   HMM   tfbs   806    813    .   +   .   M00926
human   HMM   tfbs   927    934    .   +   .   M00926
human   HMM   tfbs   944    951    .   +   .   M00926
human   HMM   tfbs   991    998    .   -   .   M00926
human   HMM   tfbs   1017   1024   .   +   .   M00926
human   HMM   tfbs   1081   1088   .   -   .   M00926
human   HMM   tfbs   1115   1122   .   -   .   M00926
human   HMM   tfbs   1199   1206   .   +   .   M00926
human   HMM   tfbs   1276   1283   .   -   .   M00926
```

Abbildung 3.8: Beispiel für eine GFF-Datei, die vom Programm ausgegeben wird. Kommentarzeilen beginnen mit einem '#'-Zeichen. Parameter, die dem Programm übergeben wurden, werden in solchen Kommentarzeilen ausgegeben. Die eigentlichen Einträge entsprechen vorhergesagten TFBSs, die durch Position, Strang und Zuordnung zu einer PSSM defi niert sind.

nicht auf die im Alignment. In der siebten Spalte wird angezeigt, ob sich die Vorhersage auf dem Vorwärts- ('+') oder dem Rückwärtsstrang ('-') befindet. Der neunte Eintrag enthält die Zuordnung zu einer Gruppe, in diesem Fall die TRANSFAC®-Zugriffsnummer der PSSM.

3.4.6 Datensätze

Tabelle 3.7: Für die Vorhersagen des HMM eingesetzte PSSMs

Zugriffsnummer	TF	N
M00761	p53	20
M00789	GATA	31
M00912	C/EBP	25
M00920	E2F	17
M00926	AP-1	33
M00931	Sp1	83
M00971	Ets	18
M00976	AHR/HIF	17
M00981	CREB/ATF	18
M01031	HNF4	24
M01034	Ebox	24

Die PSSMs, die mit mehr als 15 für den Menschen annotierten TFBSs im Datensatz I vertreten sind, wurden für Vorhersagen dieser TFBSs durch das HMM verwendet. Falls mehrere PSSMs pro TF existieren, wurde die PSSM mit der höchsten Anzahl von Mensch-TFBSs in Datensatz I ausgewählt. Eine Übersicht der untersuchten PSSMs zeigt Tabelle 3.7.

3.4.7 Vergleich der Vorhersagen von MATCH™ und des Hidden-Markov-Modells

Um zu überprüfen, ob das HMM einer rein PSSM-basierten Suche nach TFBSs überlegen ist, wurden die Vorhersagen des HMM und des Programms MATCH™ (siehe 3.2.2, S. 29) auf Nukleotidebene mit der gegebenen Annotation verglichen. Da ein Nukleotid sowohl Teil einer TFBS als auch des Hintergrundes (alle Nukleotide, die nicht zu annotierten TFBSs oder Exons gehören), sowie Teil einer Vorhersage als auch keiner Vorhersage sein kann, existieren vier verschiedene Kategorien, denen ein Nukleotid zugeteilt werden kann (siehe Tabelle 3.8 und Abbildung 3.9).

Dabei wird angenommen, dass die Nukleotide des Hintergrundes keine TFBSs enthalten. Die heutige Annotation von TFBSs im Humangenom deckt allerdings nur einen Bruchteil der real existierenden TFBSs ab, und es ist daher zu erwarten, daß die Nukleotide des Hintergrundes weitere funktionelle TFBSs enthalten.

Tabelle 3.8: Kategorisierung von Vorhersagen auf Nukleotidebene

Kategorie	Beschreibung
TP	Das Nukleotid ist Teil einer annotierten TFBS und einer vorhersagten TFBS; man spricht von einer wahr positiven („true positive“, TP) Vorhersage.
FN	Das Nukleotid ist Teil einer annotierten TFBS und keiner vorhersagten TFBS; man spricht von einer falsch negativen („false negative“, FN) Vorhersage
FP	Das Nukleotid ist Teil des Hintergrundes und einer vorhersagten TFBS; man spricht von einer falsch positiven („false positive“, FP) Vorhersage.
TN	Das Nukleotid ist Teil des Hintergrundes und keiner vorhersagten TFBS; man spricht von einer wahr negativen („true negative“, TN) Vorhersage.

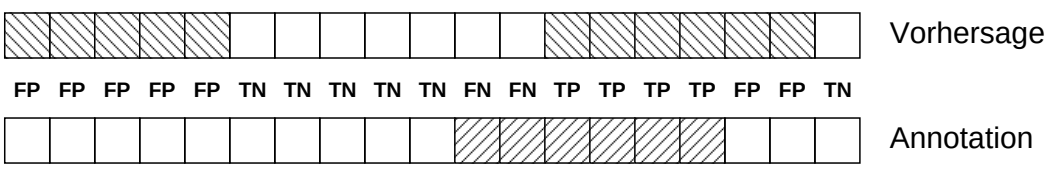

Abbildung 3.9: Vergleich von Annotation und Vorhersage auf Nukleotidebene. Jedes Kästchen repräsentiert ein Nukleotid, wobei Nukleotide, die Teil einer annotierten oder vorhergesagten TFBS sind, schraffiert hervorgehoben sind. Jedes Nukleotid fällt in eine der Kategorien TP, FN, FP oder TN.

Folgende Kenngrößen lassen sich aus den Häufigkeiten für wahr positive, falsch nega-

tive, falsch positive und wahr negative Vorhersagen berechnen:

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.24}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.25}$$

Sn steht dabei für die Sensitivität des Ansatzes und gibt an, wieviele der annotierten Nukleotide vorhergesagt werden. Der positive Vorhersagewert bzw. „positive predictive value“ PPV spiegelt den Anteil korrekt vorhergesagter Nukleotide an allen vorhergesagten Nukleotiden wider. Bei der Berechnung des positiven Vorhersagewerts ist zu beachten, dass falsch positive Vorhersagen in der Realität nicht annotierte TFBSs sein können, d.h. ein Anteil $c \in [0, 1]$ der falsch positiven Vorhersagen würde zu den wahr positiven Vorhersagen zählen. Für den realen positiven Vorhersagewert PPV_{real} würde daher gelten:

$$\begin{aligned} \text{PPV}_{\text{real}} &= \frac{\text{TP} + c \cdot \text{FP}}{\text{TP} + c \cdot \text{FP} + \text{FP} - c \cdot \text{FP}} \\ &= \frac{\text{TP} + c \cdot \text{FP}}{\text{TP} + \text{FP}} \end{aligned} \tag{3.26}$$

Der gemessene positive Vorhersagewert PPV ist daher eine Unterschätzung des realen positiven Vorhersagewertes PPV_{real}. Da dies für jede Methode zur Vorhersage von TFBSs gilt, ist der gemessene positive Vorhersagewert PPV dennoch geeignet, um diese Methoden zu vergleichen.

Das HMM wird mit verschiedenen Werten für den Übergangswahrscheinlichkeit p initialisiert, um den Wertebereich zwischen 0 und 100% Sensitivität abzudecken. Als Ausgabe erhält man eine GFF-Datei, die Positionen auf der Eingabesequenz als TFBSs klassifiziert. Diese Ausgabe wird mit der Annotation verglichen, um die Kenngrößen Sn und PPV zu berechnen.

MATCH™ ordnet Positionen auf Sequenzen einen Ähnlichkeits-Score zur PSSM, den MSS (siehe 3.2.2, S. 29), zu, der Werte zwischen 0 und 1 annehmen kann. Der MSS wird für ein Fenster der Länge λ, d.h. der Länge der PSSM, berechnet. Da der MSS für beide Stränge berechnet wird, können für ein Nukleotid bis zu 2λ Werte für den MSS berechnet werden, wobei jedem Nukleotid der maximale dieser Werte zugeordnet wird (siehe Abbildung 3.10). Setzt man nun einen Schwellenwert für den MSS fest, erhält man durch den Vergleich der Vorhersage mit der Annotation Häufigkeiten für TP, FN, FP und TN, um die

Sensitivität und den positiven Vorhersagewert zu berechnen.

```
-----------------------------------V$AP1_Q4_01(0.69)
<-----------------------------------V$AP1_Q4_01(0.03)
     ----------------------------------->V$AP1_Q4_01(0.03)
     <-----------------------------------V$AP1_Q4_01(0.27)
          ----------------------------------->V$AP1_Q4_01(0.33)
          <-----------------------------------V$AP1_Q4_01(0.03)
               ----------------------------------->V$AP1_Q4_01(0.03)
               <-----------------------------------V$AP1_Q4_01(0.92)

 C    A    A    G    T    G    A    C    T    T    C      Sequenz
0.69 0.69 0.69 0.92 0.92 0.92 0.92 0.92 0.92 0.92 0.92    MSS
```

Abbildung 3.10: Zuordnung des MSS zu Nukleotiden. Jedem Nukleotid der Beispielsequenz wird der maximale MSS der Vorhersagen, von denen es überdeckt wird, zugeordnet.

Benutzt man eine einfache Kombination aus MATCH™ und phylogenetischem Footprinting zur Vorhersage von TFBSs, so erhält man mit dem prädiktiven Ansatz (siehe 3.3.10, S. 46) für jedes Nukleotid noch einen PI-Wert, der die Sequenz-Konserviertheit widerspiegelt. Jedes Nukleotid b der Sequenz ist dabei durch ein Wertepaar $(\mathrm{MSS}, \mathrm{PI})$ charakterisiert. Diese Information kann zur Einteilung der Nukleotide in die Kategorien TP, FN, FP, TN hinzugezogen werden, indem man je einen Schwellenwert für PI-Wert und MSS setzt. Um die Wahl von zwei Schwellenwerten zu vermeiden, ist die Wahl eines einzelnen Schwellenwertes für eine Linearkombination beider Scores, der sog. lineare Score (LS) möglich:

$$\mathrm{LS} = \mathrm{PI} - m \cdot \mathrm{MSS}, \qquad m \in \mathbb{R} \tag{3.27}$$

Der Wert für m wird aus den Schwerpunkten der Wertepaare $(\mathrm{MSS}_i, \mathrm{PI}_i)$ $(i = 1, \ldots, N)$ der Nukleotide des Hintergrundes und der Wertepaare der Nukleotide $(\mathrm{MSS}_j, \mathrm{PI}_j)$ $(j = 1, \ldots, M)$ der TFBSs berechnet. Die Steigung der Senkrechten auf der Verbindungslinie zwischen diesen Schwerpunkten liefert den Wert für m:

$$m = -1 \cdot \left(\frac{\frac{1}{M}\sum_{j}^{M} \mathrm{PI}_j - \frac{1}{N}\sum_{i}^{N} \mathrm{PI}_i}{\frac{1}{M}\sum_{j}^{M} \mathrm{MSS}_j - \frac{1}{N}\sum_{i}^{N} \mathrm{MSS}_i} \right)^{-1} \tag{3.28}$$

Auch hier erhält man durch die Wahl eines Schwellenwertes für LS Häufigkeiten für TP, FN, FP und TN zur Berechnung der Kenngrößen Sn und PPV.

Kapitel 4

Ergebnisse und Diskussion

Phylogenetisches Footprinting ist ein häufig angewandter Ansatz zur Identifizierung von potentiellen TFBSs und basiert auf dem Vergleich orthologer, nicht-codierender Sequenzen. Im Rahmen dieser Arbeit wurde dieser Ansatz auf seine Gültigkeit und Nützlichkeit hin überprüft. Dazu wurde untersucht, in welchem Umfang experimentell bekannte TFBSs zwischen verschiedenen Spezies konserviert sind. Die daraus gewonnenen Erkenntnisse sollen einer verbesserten Vorhersage von unbekannten TFBSs dienen.

Die Grundvoraussetzung für erfolgreiches phylogenetisches Footprinting ist die Sicherstellung der orthologen Beziehung der zu vergleichenden Sequenzen. Diese Problematik wird in Abschnitt 4.1 behandelt. Die Konserviertheit experimentell bekannter TFBSs zwischen Mensch und Maus oder Ratte auf Sequenz- und auf Musterebene wird in den Abschnitten 4.2 und 4.3 beschrieben. Die Genregulation in Eukaryoten basiert häufig auf der Interaktion von TFs. In Abschnitt 4.4 wird die Sequenz-Konserviertheit von TFBSs, deren TFs bekannterweise miteinander interagieren, untersucht. In Abschnitt 4.5 werden orthologe Sequenzen mehrerer Spezies mit den zugehörigen Mensch-Sequenzen verglichen, da die Wahl der zu vergleichenden Spezies den Erfolg des phylogenetischen Footprintings beeinflusst. Untersucht man keine bekannten TFBSs auf ihre Konserviertheit hin, sondern möchte durch Sequenzvergleiche unbekannte TFBSs vorhersagen, benötigt man Kriterien zur Bestimmung konservierter Bereiche. In Abschnitt 4.6 wird die dafür optimale Fenstergröße zur Analyse von Alignments gesucht. Die gängigste Methode zur Vorhersage von TFBSs eines bestimmten TF beruht auf der Berechnung von Ähnlichkeiten zwischen nichtcodierenden Sequenzen und einer PSSM dieses TF. In Abschnitt 4.7 wird ein HMM beschrieben, das die Informationen einer PSSM und die Konserviertheit zwischen Spezies kombiniert, um TFBSs eines bestimmten TF vorherzusagen.

4.1 Orthologe Sequenzen

Die Voraussetzung für phylogenetisches Footprinting sind zwei (oder mehr) orthologe nicht-codierende Sequenzen. Entscheidend für den Erfolg des phylogenetischen Footprintings ist hierbei die Korrektheit der orthologen Beziehung, da sie die Grundbedingung für einen sinnvollen Vergleich von Sequenzen ist. Der intuitive Ansatz beim Vergleich orthologer Promotor-Sequenzen ist, die Sequenzen stromaufwärts der TSS bzw. in einem gewissen Abstand zur TSS in beiden Spezies zu vergleichen. Jareborg et al. (1999) und Liu et al. (2004) konzentrierten sich z.B. auf die 1000 bp, die stromaufwärts der TSS orthologer Gene in Mensch und Maus liegen. In diesem Abschnitt wird die Gültigkeit dieses Ansatzes untersucht und überprüft, inwiefern orthologe Sequenzen, basierend auf der Annotation der TSS in der Ensembl-Datenbank (siehe 3.1.4, S. 27), verlässlich bestimmt werden können.

Im Allgemeinen ist es problematisch, eine orthologe Sequenz aus einer zweiten Spezies nur anhand des bekannten Abstandes der Ausgangssequenz zur TSS im ersten Genom zu lokalisieren, denn beispielsweise kann die TSS in der zweiten Spezies an einer anderen Stelle liegen oder die TSS in einem der beiden Genome falsch annotiert sein (siehe Abbildung 4.1).

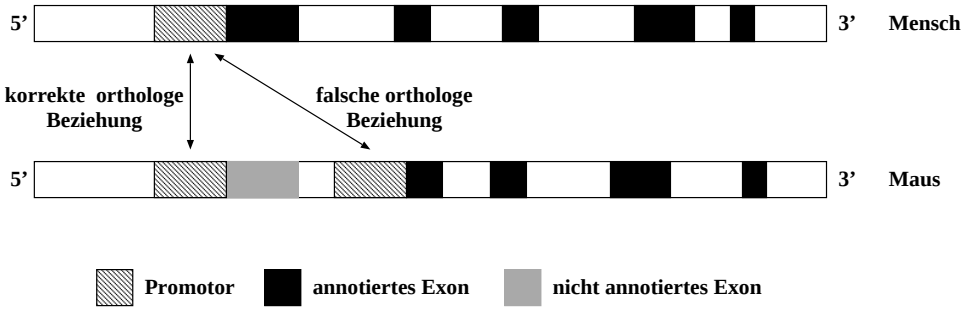

Abbildung 4.1: Bestimmung orthologer Promotoren basierend auf der Annotation der TSS. Dieses Beispiel zeigt die Exonstrukturen zweier orthologer Gene in Mensch und Maus, wobei für das Maus-Gen die Annotation des kompletten ersten Exons (grauer Kasten) fehlt. Als Promoter (gestrichelter Kasten) würde daher fälschlicherweise die Sequenz stromaufwärts des eigentlich zweiten Exons (schwarzer Kasten) angesehen werden. Der korrekte orthologe Promotor liegt aber stromaufwärts des nicht annotierten ersten Exons in der Maus.

Bei der Analyse von Enhancer-Sequenzen, die in recht großem Abstand zur TSS lokalisiert sein können, steigt zudem die Wahrscheinlichkeit für Insertionen und Deletionen zwischen der Enhancer-Sequenz und der TSS. Wenn man sich daher nur auf den relativen Abstand zur TSS verlässt, riskiert man, nicht-orthologe Sequenzen miteinander zu vergleichen. Daher benötigt man einen anderen Bezugspunkt bzw. Anker als die TSS auf dem zweiten Genom, um die korrekte orthologe Sequenz zu bestimmen. Da orthologe Se-

quenzen i.A. eine relativ hohe Ähnlichkeit zueinander aufweisen, kann ein Anker auf dem zweiten Genom definiert werden, indem dort mittels WU-BLAST (siehe 3.2.3.1, S. 32) nach hohen Sequenzhomologien zur Sequenz aus dem ersten Genom gesucht wird. Dieser Anker wird als die Sequenz aus dem zweiten Genom, die die höchste Ähnlichkeit zur Sequenz aus dem ersten Genom besitzt und gleichzeitig in der Nähe eines bekannten orthologen Gens liegt, definiert (siehe 3.3.3, S. 41). Durch dieses Verfahren ist man von der Korrektheit der Annotation der TSS unabhängig. Bei korrekter Annotation der TSS in beiden Spezies sollten die orthologen Sequenzen, die mit dieser Methode bestimmt wurden, den gleichen Abstand zur TSS haben. Ist die Annotation fehlerhaft, sollten sich die Abstände der orthologen Sequenzen zur annotierten TSS unterscheiden.

Im Folgenden wird im Detail auf die Ergebnisse der Beschaffung orthologer Sequenzen für Datensatz I (siehe 3.3.1, S. 40), der den größten Datensatz experimentell bestätigter TFBSs darstellt, eingegangen. 3383 der ursprünglichen 3508 TFBSs aus Datensatz I konnten mittels BLAT (siehe 3.2.3.3, S. 33) auf ihr entsprechendes Genom mappiert und einem Gen zugeordnet werden (siehe 3.3.2, S. 40). 72 TFBSs konnten nicht auf ihr Genom mappiert und 53 TFBSs keinem Gen zugeordnet werden konnten. Gründe dafür können noch nicht geschlossene Lücken in den genomischen Sequenzen, fehlende Annotation von Genen oder Fehler in den EMBL-Sequenzen sein. Die große Mehrheit der TFBSs ist in der Region von -500 bis +100 bp um die TSS herum lokalisiert (siehe Abbildung 4.2). 2971 (88%) der 3383 mappierten TFBSs werden von den WGAs (siehe 3.2.3, S. 30) abgedeckt, die restlichen 412 (12%) liegen in Lücken der WGAs.

Im nächsten Schritt wurde überprüft, ob für die Gene, auf deren Promotoren die 3383 TFBSs mappiert wurden, orthologe Gene annotiert sind. Für 224 der 3383 TFBSs existiert in Ensembl keine Annotation eines orthologen Gens, sodass 3159 TFBSs verblieben, für die eine WU-BLAST-Suche durchgeführt wurde. Für 481 (15%) dieser 3159 TFBSs wurde durch die WU-BLAST-Suche kein Anker auf dem zweiten Genom gefunden. Gründe dafür können wiederum Lücken in den Genomsequenzen und falsche Annotation, aber auch geringe Sequenzhomologien sein, die durch die WU-BLAST-Suche nicht entdeckt wurden. Weiterhin könnte das Maskieren von Repeats („repeat-masking“) die Ergebnisse beeinflusst haben. Dieses wurde eingeführt, um akzeptable Laufzeiten der WU-BLAST-Suche zu erhalten, falls Repeats in der Anfragesequenz enthalten sind. Aus der Literatur ist bekannt, dass regulatorische Sequenzen in repetitiver DNA vorkommen (Jordan et al., 2003). Der Repeatgehalt der Anfragesequenzen wurde daher untersucht: Repeats wurden häufiger in Anfragesequenzen gefunden, für die keine orthologe Sequenz identifiziert werden konnte (Abbildung 4.3).

Für 184 (38%) der 481 Sequenzen, für die kein Anker gefunden wurde, wurden durch-

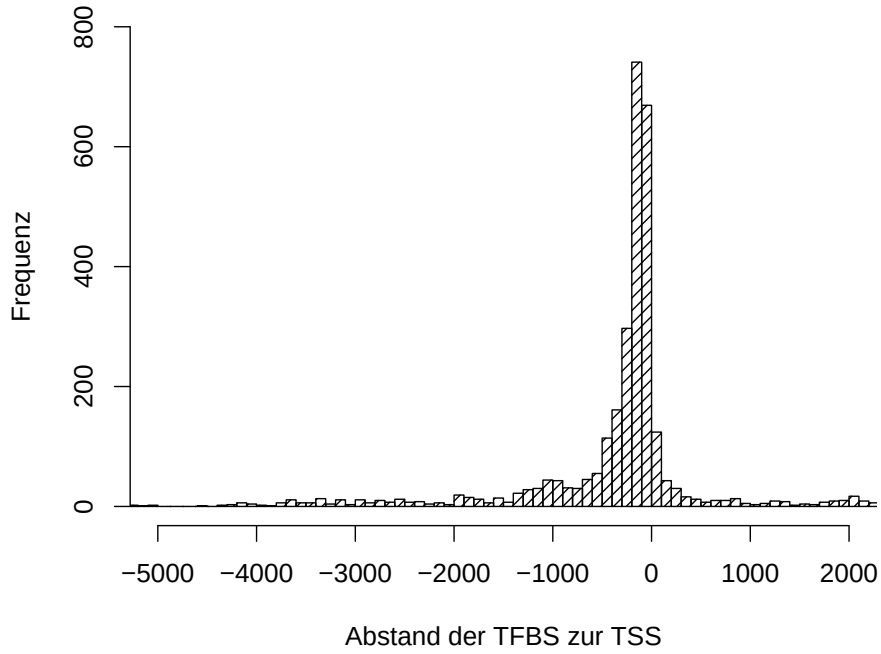

Abbildung 4.2: Die Verteilung der Abstände der in TRANSFAC® dokumentierten TFBSs zur nächst gelegenen TSS weist ein klares Maximum im Bereich von -500 bis +100 um die TSS herum auf. Dabei ist zu beachten, dass die Abstände der einzelnen TFBSs von denen der TRANSFAC®-Annotation abweichen können, da diese die Informationen in den ursprünglichen Publikationen widergibt und daher gelegentlich nicht mit den letzten Erkenntnissen über die TSS übereinstimmt.

schnittlich 34% der Basenpaare als Repeats maskiert. Für 2678 (85%) der verbleibenden 3159 TFBSs wurde eine orthologe Sequenz identifiziert. Hier war der Anteil maskierter Repeats deutlich geringer. Für 572 (21%) dieser Sequenzen wurden durchschnittlich 18% der Basenpaare maskiert. Dieser geringe Repeat-Gehalt spiegelt sich auch in den gefundenen korrespondierenden orthologen Sequenzen wider, von denen 467 (17%) Repeats mit einem durchschnittlichen Anteil von 19% der Basenpaare enthalten. Dies weist darauf hin, dass das Maskieren der Repeats in einigen Fällen die Identifizierung orthologer Sequenzen verhindert haben könnte.

Da für viele Gene in TRANSFAC® mehrere TFBSs annotiert sind und jede TFBSs in einem OSP mit einer Umgebung von 800 bp versehen wurde, überlappten viele dieser 2678 OSPs. Nach dem Zusammenfügen überlappender OSPs verblieben 614 OSPs, die 1349 Mensch und 812 Maus-TFBSs enthielten, und 161 OSPs, die 515 Ratte-TFBSs enthielten. Die Gesamtlänge der OSPs betrug ca. 733 kb, was einer Dichte von 3.65 TFBSs pro

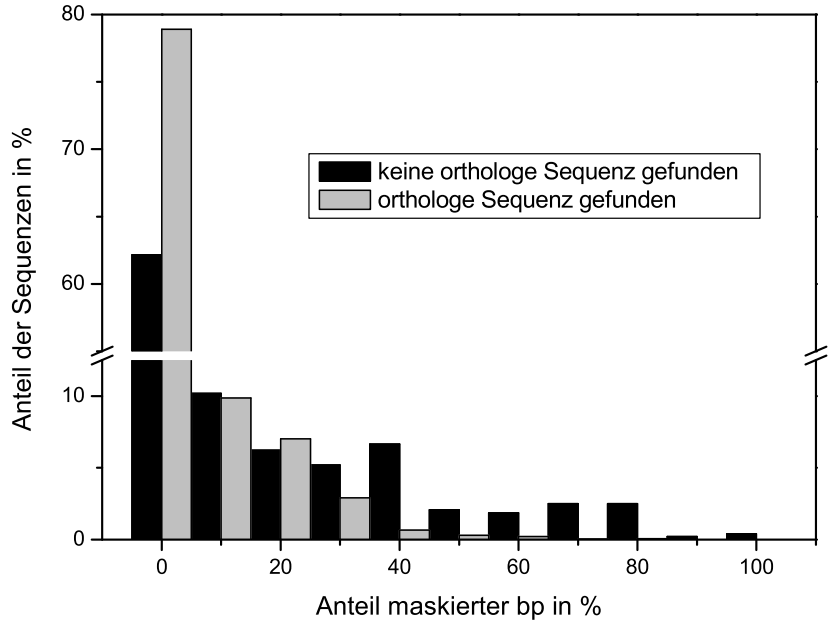

Abbildung 4.3: Verteilungen der relativen Anteile maskierter Basenpaare für die Sequenzen, bei denen ein orthologer WU-BLAST-Treffer gefunden wurde (grau, 2678 Sequenzen) oder nicht (schwarz, 481 Sequenzen).

1 kb entspricht. Für alle 775 OSPs wurden die Abstände zur TSS für die Mensch-Sequenz und die Nagetier-Sequenz verglichen (siehe Abbildung 4.4). Ein großer Unterschied dieser Abstände deutet auf ein falsche Annotation der TSS in einer oder beiden Spezies hin, d.h. die annotierten TSSs orthologer Gene wären dann keine orthologen Positionen. In 54% der Fälle unterschieden sich die Abstände zur TSS um weniger als 100 bp, aber für ungefähr 25% betrug dieser Unterschied mehr als 500 bp, für 18% sogar mehr als 1000 bp. Dies belegt, dass auf dem zweiten Genom dringend ein Anker benötigt wird, der von der annotierten TSS unabhängig ist: Da die durchschnittliche Länge der alignierten Sequenzen ca. 950 bp beträgt, würde ein Unterschied im Abstand zur TSS von 1000 bp es unmöglich machen, die bekannten TFBSs mit ihren vermuteten orthologen Gegenstücken zu alignieren, falls die orthologen Sequenzen in Bezug auf die jeweilige TSS als Anker beschafft worden wären.

Eine ähnliche Beobachtung machten auch Prakash und Tompa (2005) in ihrer Studie. Sie kommen zu dem Schluss, dass die annotierten Translationsstartstellen (TLSs) orthologer Gene oft keine orthologen Positionen sind. Dazu wurden in WGAs zwischen Mensch und Maus die Nukleotide in der Maus bestimmt, die zu einer TLS im Menschen aligniert sind. Für diese Nukleotide wurde der Abstand zur annotierten TLS des orthologen Maus-

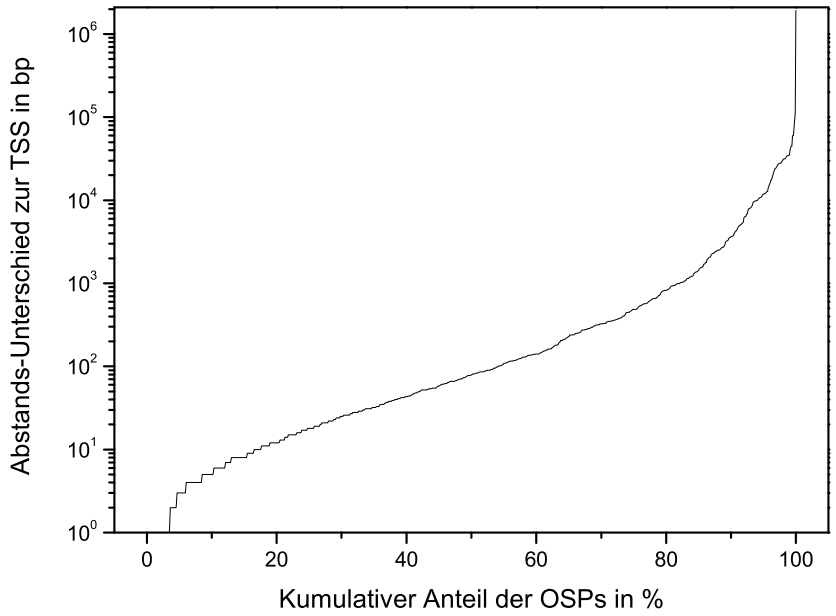

Abbildung 4.4: Der Unterschied im Abstand zur TSS für die 775 OSPs ist gegen den kumulativen Anteil der OSPs aufgetragen, die maximal diesen Abstands-Unterschied aufweisen. Ungefähr 25% der OSPs unterscheiden sich um mehr als 500 bp in ihrem Abstand zur TSS.

Gens bestimmt. Für 25% der Maus-Gene ist dieser Abstand größer als 1000 bp, was vermuten lässt, dass die ersten codierenden Exons in diesen Fällen nicht ortholog sind. Gründe dafür können der Verlust des ersten codierenden Exons in einer Spezies oder fehlerhafte Annotation sein.

Zusammenfassend lässt sich sagen, dass die Korrektheit der orthologen Beziehung zwischen Sequenzen sichergestellt werden muss, da der erste Schritt beim phylogenetischen Footprinting die Beschaffung orthologer Promotoren ist und der Erfolg weiterer Schritte entscheidend von der Korrektheit der orthologen Beziehung der Promotoren abhängt.

4.2 Sequenz-Konserviertheit von Transkriptionsfaktor-Bindestellen

Der Grundgedanke des phylogenetischen Footprintings ist, dass sich in einem Alignment orthologer nicht-codierender Sequenzen TFBSs von nicht-funktionellen Bereichen durch eine höhere Konserviertheit abheben. Ein Ziel dieser Arbeit war u.A. zu überprüfen, welches Alignment-Programm und welche Definition der Konserviertheit am besten geeignet ist, um TFBSs zu identifizieren. Dazu wurden mehrere verschiedene Alignment-Algorithmen (siehe 3.2.3, S. 30) verglichen und der Einfluss des KS (siehe 3.3.5, S. 43) untersucht. Die erhaltenen Ergebnisse sollten Hinweise darauf geben, wie man phylogenetisches Footprinting optimal zur Identifizierung unbekannter TFBSs durch Mensch-Nagetier-Vergleiche anwenden kann. Zwei Parameter sind dazu von entscheidender Bedeutung: die Sensitivität und die Spezifität. Der Begriff Sensitivität ist allerdings irreführend, da nicht alle TFBSs zwischen Mensch und Nagetieren konserviert sind, und damit eine Sensitivität von 100% nicht erreicht werden kann. Um einem Missverständnis vorzubeugen, wird daher im Folgenden stattdessen der Begriff „Konserviertheitsrate“ verwendet. Die Konserviertheitsrate C_{seq} kann relativ einfach als der Anteil der konservierten TFBSs an allen untersuchten TFBSs berechnet werden (siehe 3.3.5, S. 43).

Viel schwieriger ist es dagegen, die Spezifität zu bestimmen, da nicht-funktionelle Regionen, die als konserviert identifiziert wurden (falsch positives Ergebnis), nicht von potentiell funktionellen, konservierten Regionen unterschieden werden können. Um abzuschätzen, wie hoch der Anteil falsch positiver Ergebnisse ist, wurden alle Hintergrund-Sequenzen (siehe 3.3.5, S. 44) auf ihre Konserviertheit hin untersucht. Der Anteil konservierter Hintergrund-Sequenzen an allen Hintergrund-Sequenzen ergibt die Hintergrund-Konserviertheitsrate C_{seq}^{bg} (siehe 3.3.5, S. 44). Die Hintergrund-Konserviertheitsrate ist allerdings eine Überschätzung des Anteils falsch positiver Ergebnisse, da nur ein Bruchteil aller TFBSs annotiert ist und andere, noch nicht annotierte funktionelle TFBSs in der Umgebung der bekannten TFBSs existieren sollten. Wenn diese unbekannten TFBSs konserviert wären, würden sie als falsch positive Ergebnisse gezählt werden, obwohl sie in Wirklichkeit wahr positive Ergebnisse wären. Die Hintergrund-Konserviertheitsrate C_{seq}^{bg} ist daher eine sehr konservative Einschätzung des Anteils falsch positiver Ergebnisse.

Das Problem bei der Abschätzung des Anteils falsch positiver Ergebnisse an allen Vorhersagen ist daher, dass nicht genau bekannt ist, welche Bereiche des Genoms nicht-funktionell sind. Eine Möglichkeit, dies zu umgehen, ist, nicht-funktionelle Bereiche artifiziell zu erzeugen. Um ein zweites Maß für den Anteil falsch positiver Ergebnisse zu bekommen, wurden OSPs, die randomisierte TFBSs enthalten, aligniert (siehe 3.3.6, S. 44).

Der Anteil konservierter randomisierter TFBSs an allen randomisierten TFBSs ergibt die Shuffle-Konserviertheitsrate C_{seq}^{shuf} (siehe 3.3.6, S. 44). Da die flankierenden Bereiche auf beiden Seiten einer TFBS immer noch alignierbar sind, oder anders gesagt, die orthologe Beziehung zwischen beiden Sequenzen noch besteht, wurde die TFBS durch die flankierenden Sequenzen in ein Alignment mit dem Abschnitt der orthologen Sequenz, zu dem die nicht randomisierte TFBS aligniert wurde, gezwungen. Auf diese Weise wurde empirisch bestimmt, inwiefern die Nukleotid-Zusammensetzung der TFBSs deren PI-Werte im Alignment beeinflusst. Die Shuffle-Konserviertheitsrate C_{seq}^{shuf} spiegelt daher die Wahrscheinlichkeit wider, bei gegebener Nukleotid-Zusammensetzung der TFBS einen bestimmten PI-Wert für diese zufällig zu erhalten.

Abbildung 4.5 zeigt die unterschiedlichen Verteilungen der PI-Werte für Alignments der OSPs aus Datensatz I mit AVID. Die randomisierten TFBSs weisen eine geringere Sequenz-Konserviertheit (durchschnittlicher PI-Wert von 42.6%) als die Hintergrund-Sequenzen (52.0%) und die bekannten TFBSs (75.6%) auf.

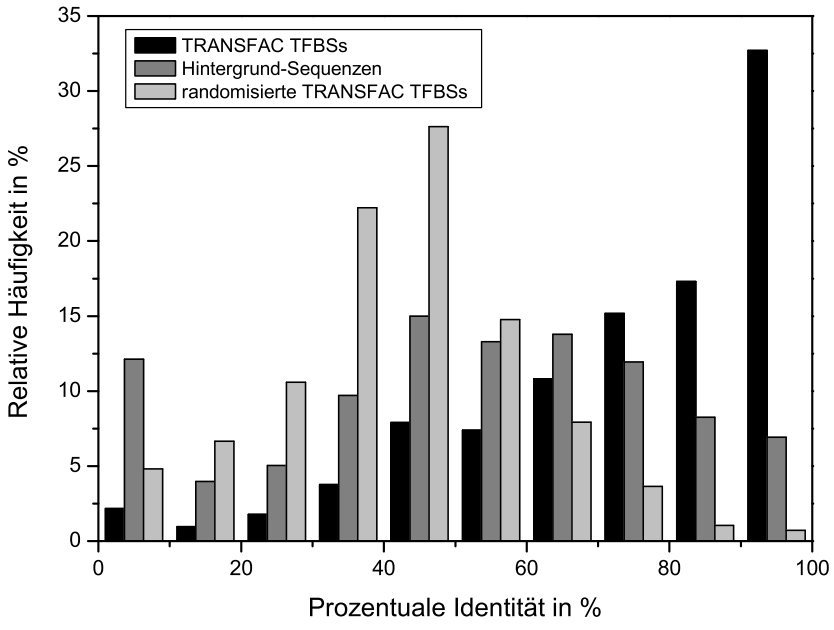

Abbildung 4.5: PI-Wert-Verteilungen der TFBSs (schwarz), der Hintergrund-Sequenzen (dunkelgrau) und der randomisierten TFBSs (hellgrau). Die gezeigten Verteilungen basieren auf Alignments, die mit AVID für die OSPs aus Datensatz I angefertigt wurden.

Der relative Anteil konservierter TFBSs, Hintergrund-Sequenzen sowie randomisierter TFBSs ist abhängig vom gewählten KS (siehe Abbildung 4.6). Die TFBSs sind für alle KSs

stärker konserviert als die Hintergrund-Sequenzen und die randomisierten TFBSs. Dies zeigt, dass der Ansatz des phylogenetischen Footprintings für Mensch-Nagetier-Vergleiche in der Lage ist, funktionelle TFBSs von der umgebenden Sequenz abzuheben. Die Shuffle-Konserviertheitsrate C_{seq}^{shuf} sinkt bis zu einem KS von ca. 65% stark ab und erreicht für einen KS von 100% einen sehr niedrigen Wert von 0.7%. Im Gegensatz dazu sind 24.9% der TFBSs und 5.1% der Hintergrund-Sequenzen vollständig konserviert.

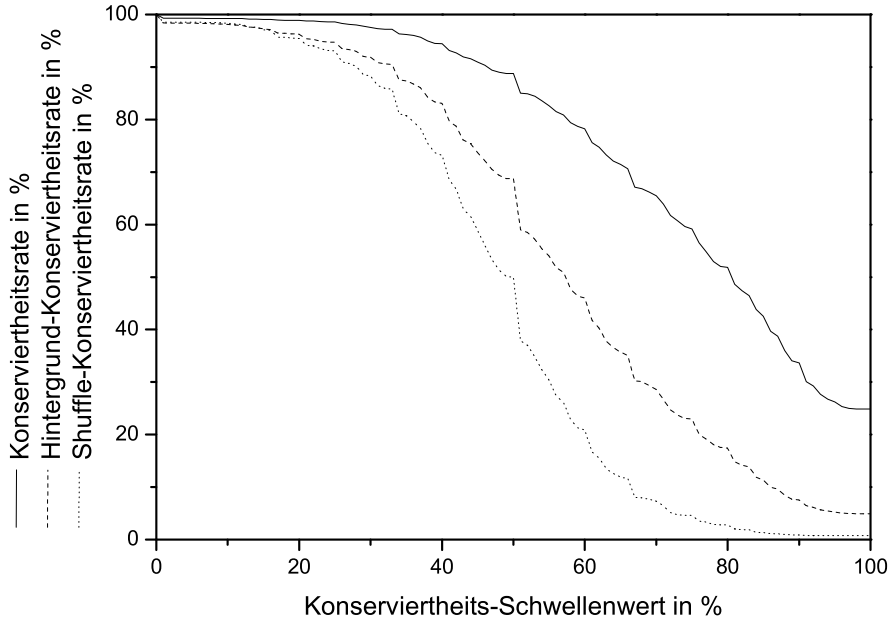

Abbildung 4.6: Abhängigkeit der Konserviertheitsraten vom KS. Die Konserviertheitsraten für die TFBSs, die Hintergrund-Sequenzen und die randomisierten TFBSs zeigen eine unterschiedliche Abhängigkeit vom KS. Die TFBSs sind dabei für alle Werte des KS stärker als der Hintergrund und die randomisierten TFBSs konserviert. Die gezeigten Konserviertheitsraten basieren auf Alignments, die mit AVID für die OSPs aus Datensatz I angefertigt wurden.

Einfluss des Alignment-Programms

Um verschiedene Alignment-Algorithmen miteinander zu vergleichen, wurden für jedes Alignment-Programm die erhaltenen Konserviertheitsraten C_{seq} gegen die korrespondierenden Hintergrund-Konserviertheitsraten C_{seq}^{bg} und Shuffle-Konserviertheitsraten C_{seq}^{shuf} in einer sogenannten „receiver operating characteristic“-Kurve (ROC-Kurve) aufgetragen (siehe Abbildung 4.7).

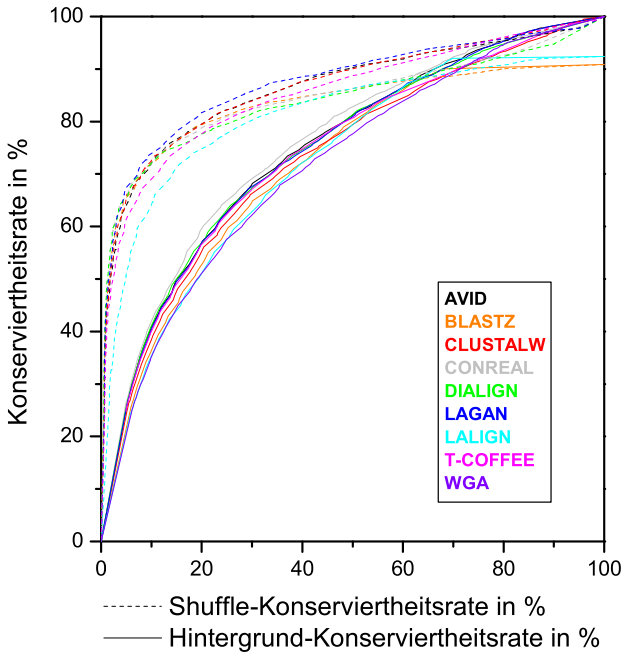

Abbildung 4.7: ROC-Diagramm für verschiedene Alignment-Algorithmen. Zum Vergleich der Performanz der unterschiedlichen Alignment-Algorithmen wurden die erhaltenen Konserviertheitsraten gegen die Hintergrund- und Shuffle-Konserviertheitsraten aufgetragen. Die globalen Alignment-Programme trennen die TFBSs geringfügig besser von nicht-funktionellen Bereichen ab als die lokalen Alignment-Programme und die WGAs.

Dabei zeigten sich keine signifikanten Unterschiede zwischen den untersuchten Alignment-Algorithmen, wobei globale Alignment-Programme eine geringfügig bessere Performanz als die lokalen Alignment-Programme und die BLASTZ-WGAs lieferten. Die von lokalen Tools erzeugten Alignments decken die Eingabesequenzen nicht vollständig ab, was zu einer verringerten Sensitivität bei der Detektion von TFBSs führt. Dies wurde auch von Pollard et al. (2004) beobachtet, die in ihrer Studie Alignments über eine Spanne von evolutionären Abständen simulierten und diese mit den Alignments verglichen, die von paarweisen Alignment-Programmen produziert wurden.

Obwohl die BLASTZ-WGAs eine etwas schwächere Performanz als die anderen Alignment-Programme aufwiesen, deckten sie einen größeren Anteil der annotierten TFBSs ab, da 2971 TFBS auf die WGAs mappiert werden konnten, wohingegen nur 2678 TFBSs von den OSPs, die mittels WU-BLAST erhalten wurden, abgedeckt wurden. Die höhere Abdeckung der TFBSs aus Datensatz I durch die WGAs ist darauf zurückzuführen, dass hier weder die Kenntnis eines annotierten orthologen Gens noch eine erfolgreiche WU-BLAST-Suche nötig ist. Die schwächere Performanz kann auf den gleichen Grund zurückgeführt

werden, da für die WGAs nicht sichergestellt wurde, dass orthologe Sequenzen aligniert wurden.

CONREAL trennte im Vergleich mit den anderen Programmen die bekannten TFBSs am besten von nicht-funktionellen Sequenzen ab. Dabei muss aber betont werden, dass CONREAL im Gegensatz zu den anderen Alignment-Programmen bekannte TFBSs in Form von PSSMs zur Erzeugung des Alignments benutzt (siehe 3.2.3, 30). Die Zirkularität dieses Ansatzes mag ein Nachteil sein, da Sequenzabschnitte, die keinem bekannten TF zugeordnet werden können, nicht aligniert werden. Dieser Nachteil gilt nicht für die anderen Alignment-Algorithmen, deren Laufzeiten zusätzlich mindestens um das 20fache kürzer sind als die von CONREAL. Für alle weiteren Untersuchungen von Datensatz I wurden mit AVID berechnete Alignments verwendet.

Die geringen Performanz-Unterschiede zwischen den einzelnen Alignment-Programmen sind dadurch zu erklären, dass Mensch- und Nagetier-Sequenzen eine genügend hohe Ähnlichkeit aufweisen, sodass die meisten Alignment-Programme ähnliche Ergebnisse produzieren. Die Performanz-Unterschiede können jedoch für größere evolutionäre Abstände deutlich stärker sein, wie von Pollard et al. (2004) gezeigt wurde. Sehr hohe evolutionäre Abstände können es sogar unmöglich machen, ein genaues Alignment zu erhalten, wie von Rosenberg (2005) beschrieben wurde. In dieser Studie wurde gezeigt, dass, wenn für zwei nicht-codierende Sequenzen weniger als 50% der Positionen identisch sind, CLUSTALW paarweise Alignments produziert, die sich nicht von denen zufällig erzeugter Sequenzen unterscheiden. Daher ist für Sequenz-Vergleiche zwischen Spezies, die entfernter verwandt sind als Mensch und Nagetier, anzunehmen, dass ein bestimmtes Alignment-Programm den anderen vorzuziehen ist.

Bestimmung des optimalen KS

Um den KS zu bestimmen, der einen möglichst hohen Wert für C_{seq} und gleichzeitig möglichst niedrige Werte für C_{seq}^{bg} bzw. C_{seq}^{shuf} ergibt, wurde der Abstand jedes Punktes im ROC-Diagramm zur linken oberen Ecke des Diagramms berechnet (siehe Abbildung 4.8).

An diesem Punkt beträgt die Konserviertheitsrate für die bekannten TFBSs 100%, während der Anteil falsch positiver Vorhersagen 0% ist. Ein minimaler Abstand zu diesem optimalen Punkt ist für einen KS von 65% gegeben. Dieser Schwellenwert ergibt die beste Diskrimination zwischen den realen TFBSs und sowohl den Hintergrund-Sequenzen als auch den randomisierten TFBSs, und wird daher für nachfolgende Analysen verwendet. Für diesen KS erhält man eine Konserviertheitsrate C_{seq} von 71.7%, eine Hintergrund-Konserviertheitsrate C_{seq}^{bg} von 35.2% und eine Shuffle-Konserviertheitsrate C_{seq}^{shuf} von 9.8%. Der Schwellenwert von 65% wurde für die Gesamtheit aller TFBSs in Datensatz I berech-

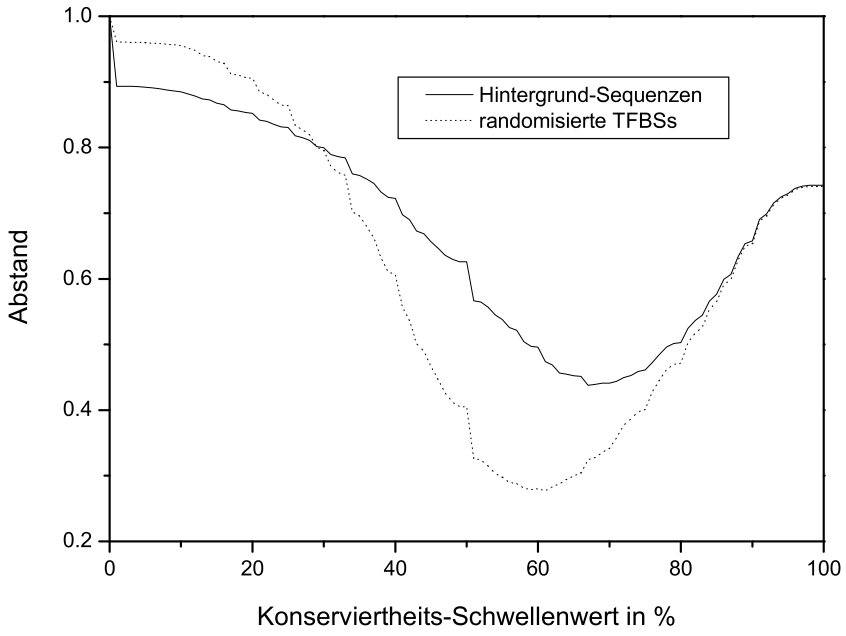

Abbildung 4.8: Für jeden KS wurde der Abstand zur linken oberen Ecke des ROC-Diagramms berechnet (für AVID-Alignments). Je kleiner der Abstand, um so besser ist der Kompromiss zwischen der Konserviertheitsrate und falsch positiven Vorhersagen. Für einen KS von 65% erreicht man eine optimale Trennung sowohl zwischen den TFBSs und den Hintergrund-Sequenzen als auch zwischen den TFBSs und den randomisierten TFBSs.

net, für einzelne TFs ist es allerdings denkbar, dass der optimale KS leicht abweicht.

Mit diesem Schwellenwert verbleiben 28.3% der TFBSs, die als nicht konserviert angesehen werden. Ein Teil dieser TFBSs ist möglicherweise Spezies-spezifisch, d.h. nur im Menschen oder in den Nagetieren aktiv, und daher durch Sequenzvergleiche nicht detektierbar. Weiterhin ist es denkbar, dass regulatorische Bereiche aufgrund ihrer inneren Variabilität TFBSs verloren oder hinzugewonnen haben, oder dass sich deren Reihenfolge und/oder Abstände verändert haben. Letzteres konnte z.B. für *Drosophila* von Ludwig et al. (2000) gezeigt werden.

Eine Teilmenge von 100 Nagetier-TFBSs wurde zu Mensch-TFBSs aligniert, an die der gleiche TF bindet, d.h. in diesen Fällen sind in der TRANSFAC®-Datenbank sowohl die Mensch-TFBS als auch ihr orthologes Gegenstück annotiert. 94 dieser TFBSs haben einen PI-Wert $\geq 65\%$. Dies ist ein eindeutiger Hinweis darauf, dass TFBSs, die in beiden Spezies gleichermaßen existieren, mittels phylogenetischem Footprinting sehr gut detektiert werden können. Die 100 Nagetier-TFBSs wurden nicht zur Berechnung der Konserviert-

heitsrate C_{seq} verwendet, um keine Informationen doppelt zu zählen.

Da für 481 TFBSs keine orthologe Sequenz gefunden wurde, könnte die obige Konserviertheitsrate C_{seq} von 71.7% eine Überschätzung der realen Konserviertheitsrate sein. Wenn die Beschaffung orthologer Sequenzen auf Grund einer insgesamt geringen Sequenzhomologie gescheitert ist, wären diese TFBSs mit einer geringeren Wahrscheinlichkeit konserviert und würden die Prozentzahl konservierter TFBSs verringern. In einem „worst case"-Szenario, d.h. unter der Annahme, dass keine dieser 481 TFBSs konserviert ist, ergibt sich eine Konserviertheitsrate C_{seq} von 60.4% als eine Minimalschätzung der genomweiten Konserviertheit von TFBSs zwischen Mensch und Nagetieren.

Vergleich mit früheren Studien

Der in dieser Arbeit beobachtete Anteil von 72% konservierter TFBSs bestätigt die Ergebnisse früherer Studien. Levy und Hannenhalli (2002) fanden 65% von 485 in der TRANSFAC®-Datenbank (Version 4.4) annotierten TFBSs in Bereichen, die zwischen Mensch und Maus konserviert waren. Konservierte Bereiche wurden dabei über eine minimale Länge von 50 bp und einen minimalen PI-Wert von 70% bestimmt. Lenhard et al. (2003) detektierten ca. 68% von 150 bekannten TFBSs (40 manuell annotierte TFBSs und 110 TFBSs aus der TRANSFAC®-Datenbank, Version 4.0) in konservierten Regionen, die über ein „sliding window" der Länge 50 bp und einen KS von 70% definiert wurden. Liu et al. (2004) benutzten einen KS von 70% und eine Fenstergröße von 21 bp, die auf 330 bekannten TFBSs (TRANSFAC®-Datenbank, Version 6.2) zentriert wurde. Mit diesen Parametern wurden 60% der TFBSs und ca. 25% der Hintergrund-Sequenzen als konserviert eingestuft.

Für die in diesem Abschnitt beschriebenen Untersuchungen wurde nicht wie beim prädikitven Ansatz (siehe 3.3.10, S. 46) eine bestimmte Fenstergröße zur Analyse benutzt, sondern es wurden die TFBSs selbst untersucht. Dermitzakis und Clark (2002) verfolgten den gleichen Ansatz und in ihrer Studie hatten ca. 62% von 64 experimentell bekannten TFBSs, die der Primärliteratur und der TRANSFAC®-Datenbank entnommen wurden, einen PI-Wert größer oder gleich einem KS von 70%. Für den gleichen KS sind 66% der TFBSs und 28% der Hintergrund-Sequenzen in Datensatz I konserviert.

Es ist anzumerken, dass in den obigen Studien, die eine Fenstergröße von 50 bp benutzten, die Konserviertheitsrate der TFBSs größer ist als in den Studien, die mit kleineren Fenstergrößen arbeiteten. Eine Erklärung dafür könnte sein, dass die TFBSs innerhalb eines Fensters der Länge 50 bp auch PI-Werte $\leq 70\%$ besitzen können, da die durchschnittliche Länge einer TFBS viel kleiner als 50 bp ist.

Sequenz-Konserviertheit der „Regulatory Features“der cisRED-Datenbank

Der Ensembl Release 34 (5. Oktober 2005) enthält die sogenannten „regulatory features“ (RFs) der cisRED-Datenbank (Robertson et al., 2006) (Version 1.2e, http://www.cisred.org, siehe auch 3.1.3, S. 27). Im Gegensatz zu den annotierten TFBSs in TRANSFAC® sind die RFs vorhergesagte TFBSs, d.h. es gibt keinen experimentellen Beleg dafür, dass diese Sequenzen funktionell sind. Um die Qualität dieser vorhergesagten TFBSs einschätzen zu können, wurden sie auf ihre Konserviertheit hin untersucht. Dabei ist zu erwarten, dass der Anteil konservierter RFs geringer ist als der Anteil konservierter TRANSFAC®-TFBSs.

Repräsentativ für die Daten der cisRED-Datenbank wurden alle RFs, die auf dem Chromosom 1 des Menschen liegen (Datensatz II), der gleichen Analyse wie die TRANSFAC®-TFBSs aus Datensatz I unterzogen, d.h. die RFs wurden Genen zugeordnet (siehe 3.3.2, S. 40) und zugehörige orthologe Sequenzen identifiziert, wobei überlappende OSPs zusammengefasst wurden (siehe 3.3.3, S. 41). Es wurden 20332 RFs in 732 OSPs erhalten, die eine Gesamtlänge von ca. 1.7 Mb haben. Dies entspricht einer Dichte von 11.83 RFs pro 1 kb. Diese Dichte ist dreimal so hoch wie die in Datensatz I, welcher 3.65 TFBSs pro 1 kb enthält (Gesamtlänge ca. 733 kb). Die 732 OSPs wurden mit AVID aligniert.

Die Konserviertheit der RFs ist für alle Werte des KS größer als die der Hintergrund-Sequenzen (siehe Abbildung 4.9A), was ein Hinweis auf eine vorhandene Funktionalität innerhalb der RFs ist, wobei dieser Unterschied allerdings gering ist: Für einen KS von 65% sind 36.5% der RFs und 28.8% der Hintergrund-Sequenzen konserviert. Im Vergleich dazu sind 71.7% und 35.2% der TFBSs aus Datensatz I konserviert. Die höhere Hintergrund-Konserviertheitsrate der TRANSFAC®-TFBSs ist auf die kürzeren OSPs (durchschnittliche Länge von 950 bp) zurückzuführen, wobei durschnittlich 11% der Alignmentpositionen Lücken sind. Die OSPs aus Datensatz II haben hingegen eine durchschnittliche Länge von 2.4 kb und reichen dadurch stromaufwärts weiter von der TSS weg. Die in größerer Entfernung zur TSS gelegenen Bereiche weisen eine geringere Homologie auf und führen damit zu einem relativ hohen Anteil schlecht konservierter Hintergrund-Sequenzen in Datensatz II. Der relative Anteil an Lücken in den Alignments beträgt hier ca. 24%. Ignoriert man zur Berechnung der Konserviertheitsraten alle PI-Werte, die gleich 0% sind, so nähert sich die Hintergrund-Konserviertheitsrate der cisRED-RFs denen der TRANSFAC®-TFBSs an (siehe Abbildung 4.9B).

Im ROC-Diagramm (siehe Abbildung 4.10) weicht die Kurve für die RFs im Gegensatz zu den TRANSFAC®-TFBSs nur schwach von der Diagonalen ab. Die im Vergleich zu den TRANSFAC®-TFBSs viel schwächere Konserviertheit deutet darauf hin, dass nur ein kleiner Anteil der RFs funktionell ist.

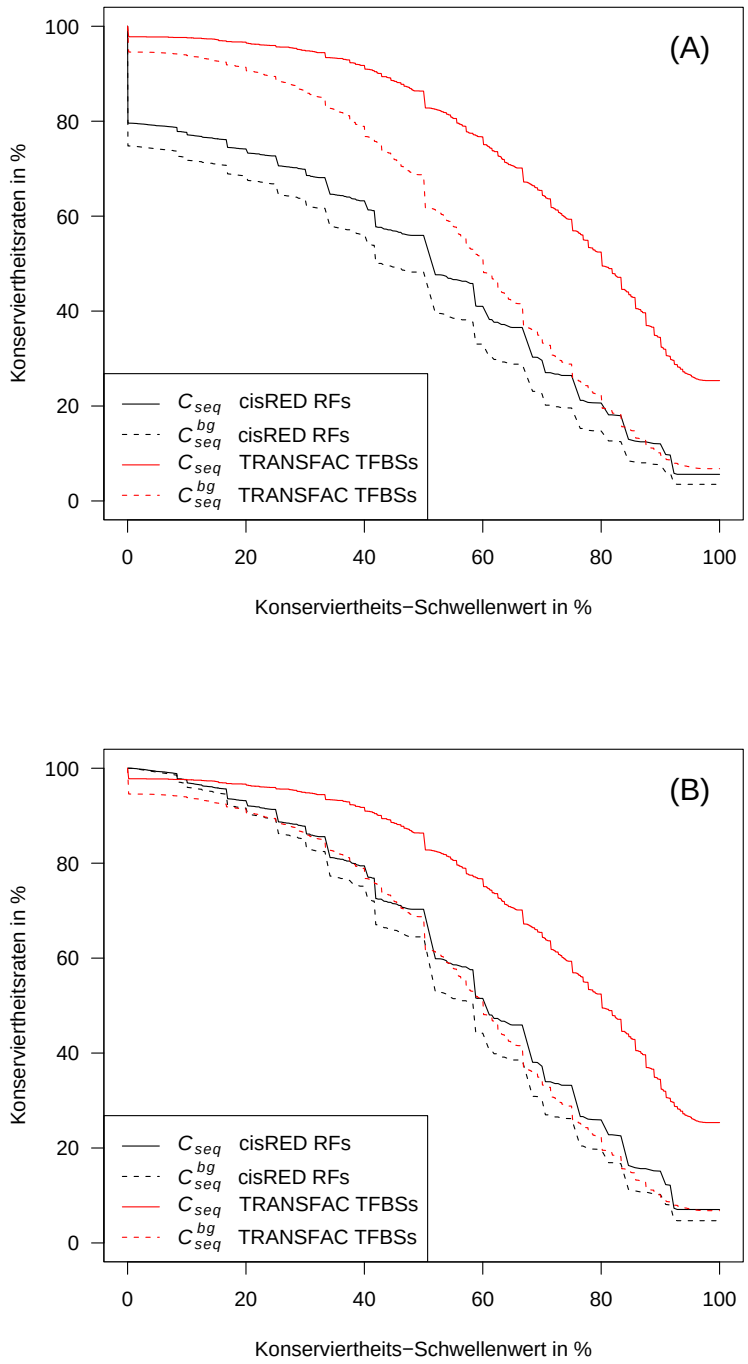

Abbildung 4.9: (A) Die RFs der cisRED-Datenbank sind für alle Werte des KS stärker konserviert als die Hintergrund-Sequenzen, jedoch ist dieser Unterschied sehr gering: Für einen KS von 65% sind beispielsweise 36.5% der RFs und 28.8% der Hintergrund-Sequenzen konserviert. Dieser Unterschied ist für die TRANSFAC®-TFBSs deutlich höher. **(B)** Es wurden alle zu 0% konservierten Sequenzen von der Berechnung der Konserviertheitsraten für die cisRED-RFs ausgeschlossen. Dadurch nähert sich die Hintergrund-Konserviertheitsrate der cisRED-RFs der der TRANSFAC®-TFBSs an.

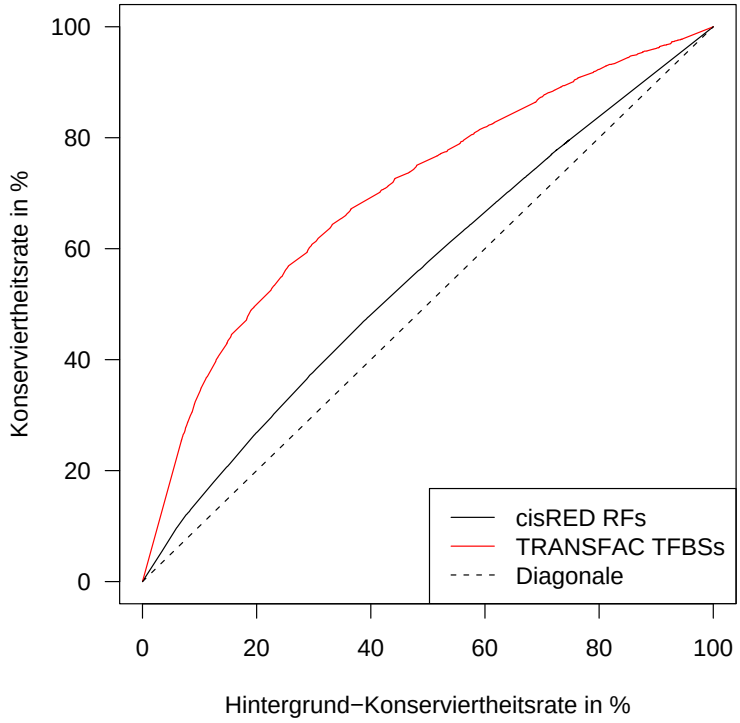

Abbildung 4.10: Die ROC-Kurve für 20332 RFs auf dem Chromosom 1 des Menschen weicht leicht von der Diagonalen ab, d.h. die RFs sind etwas stärker konserviert als die Hintergrund-Sequenzen. Die TRANSFAC®-TFBSs heben sich deutlicher von den Hintergrund-Sequenzen ab.

4.2.1 Abhängigkeit der Konserviertheitsrate vom Transkriptionsfaktor

Die untersuchten TFBSs aus Datensatz I wurden bisher in ihrer Gesamtheit auf ihre Konserviertheit hin untersucht. Da TFBSs aber von unterschiedlichen TFs gebunden werden, wurden die TFBSs anhand ihrer zugehörigen TFs gruppiert und die Konserviertheitsrate für jede dieser Untergruppen einzeln bestimmt. Dabei wurden deutliche Unterschiede festgestellt (siehe Tabelle 4.1).

Tabelle 4.1: Sequenz-Konserviertheitsrate C_{seq} für TFBSs bestimmter TFs (für AVID-Alignments der OSPs aus Datensatz I)

TF	N*	C_{seq}†	p_{seq}‡
MyoD	16	100.0%	5.3E-06
MEF-2A	13	100.0%	0.00012
SRF	33	96.9%	6.4E-12
NF-AT1	31	96.7%	4.9E-11
USF-1	22	86.3%	0.0001
NF-κB	38	84.2%	6.7E-08
CREB	49	83.6%	3.8E-10
c-Ets-1	18	83.3%	0.00368
ER-α	17	82.3%	0.00876
MITF	24	79.1%	0.00128
C/EBPβ	42	78.5%	1.1E-06
Sp3	27	77.7%	0.00073
RXR-α	54	75.9%	1.2E-07
HNF-1α	20	75.0%	0.03155
AP-1	77	74.0%	4.2E-10
YY1	29	72.4%	0.00469
POU2F1	32	65.6%	0.04194
Sp1	229	63.7%	1.5E-16
C/EBPα	87	63.2%	9.3E-06
c-Myb	16	62.5%	1
p53	27	51.8%	1
Nkx2-1	33	51.5%	1
AP-2αA	29	48.2%	1
GATA-1	50	44.0%	1

*Anzahl der TFBSs pro TF.

†Die Hintergrund-Konserviertheitsrate C_{seq}^{bg} beträgt 35.22%.

‡Die Wahrscheinlicheit p, die beobachtete (oder eine größere) Differenz zwischen C_{seq} und C_{seq}^{bg} zufällig zu erhalten, wurde mit Gleichung (3.7) berechnet. Die p-Werte wurden Bonferroni-korrigiert.

TFBSs von MyoD ($C_{seq} = 100\%$), MEF-2 (100%), SRF (96.9%) oder NF-AT1 (96.7%) haben höhere Konserviertheitsraten auf Sequenzebene als TFBSs für Sp1 (63.7%), C/EBPα (63.2%), AP-2αA (48.2%) oder GATA-1 (44.0%). Diese Ergebnisse sind in guter Übereinstimmung mit den Beobachtungen von Wasserman et al. (2000), die einen kleineren Datensatz für vier verschiedene TFs untersuchten. In ihrer Studie waren alle 20 MEF2, 23 MyoD und 15 SRF TFBSs konserviert, jedoch nur 18 von 24 Sp1 TFBSs.

Um zu überprüfen, ob sich die beobachteten Werte für C_{seq} signifikant von der Hintergrund-Konserviertheit abheben, wurde mit Gleichung (3.7) (siehe 3.3.9, S. 46) die Wahrscheinlichkeit berechnet, den beobachteten Wert für C_{seq} zufällig zu erhalten. Die berechneten Wahrscheinlichkeiten wurden mit der konservativen Bonferroni-Methode (Bland und Altman, 1995) für multiples Testen korrigiert. Die beobachteten Werte für die Konserviertheitsrate C_{seq} einiger TFs wie z.B. AP-2αA, p53 und GATA-1 unterschieden sich nicht signifikant von der Hintergrund-Konserviertheit, weshalb die Sequenz-Konserviertheit als alleiniges Kriterium nicht ausreicht, um TFBSs für diese TFs zu identifizieren.

Die TFs, die eine hohe Konserviertheitsrate ihrer TFBSs aufweisen, spielen möglicherweise sowohl für den Menschen als auch für die Nagetiere eine entscheidende Rolle in der Genregulation, weshalb der selektive Druck auf ihre TFBSs erhöht ist. Für TFs mit einer niedrigen Konserviertheitsrate ihrer TFBSs ist es denkbar, dass diese TFs sich durch eine sehr tolerante Sequenzerkennung auszeichnen und daher ein geringerer evolutionärer Druck auf den zugehörigen TFBSs lastet als auf TFBSs, an die TFs mit einem sehr strikten Bindungsprofil binden. Diese Hypothese wird allerdings dadurch widerlegt, daß es keine Korrelation (Spearman'scher Korrelationskoeffizient ρ von 0.092 bei einem zugehörigen p-Wert von 0.417; zur Berechnung siehe Press et al., 1992) zwischen dem durchschnittlichen Informationsgehalt von PSSMs und den Konserviertheitsraten der zugehörigen TFBSs gibt (siehe Abbildung 4.11).

Eine weitere denkbare Erklärung für niedrige Konserviertheitsraten der TFBSs bestimmter TFs ist, dass die nicht konservierten TFBSs in einem homotypischen Cluster im Genom eines gemeinsamen Vorfahrens existierten, wobei unterschiedliche Elemente in den einzelnen evolutionären Linien mutiert sind und ihre Funktionen verloren haben. Die verbleibenden Elemente des homotypischen Clusters haben diesen Funktionsverlust möglicherweise kompensiert. Ein Beispiel dafür ist der *even-skipped* stripe 2 Enhancer (S2E) in *Drososphila*, der die Expression des Gens *even-skipped (eve)* reguliert. Ludwig et al. (2000) beschrieben, dass die Expressionsmuster von *eve* in verschiedenen Spezies nahezu identisch sind, obwohl die S2E-Sequenzen erheblich divergierten. Die Autoren zeigten, dass diese Unterschiede in den Sequenzen zwar funktionelle Konsequenzen haben, welche wiederum aber durch andere Unterschiede ausgeglichen werden. Für einige der TFs

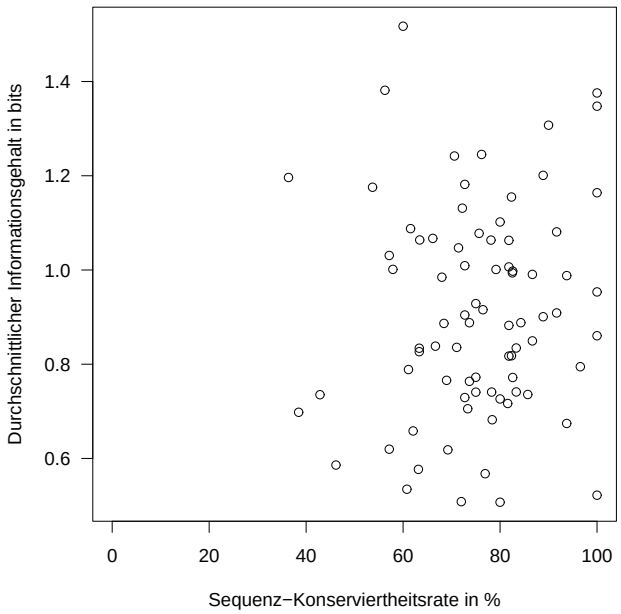

Abbildung 4.11: Durchschnittlicher Informationsgehalt von TRANSFAC®-PSSMs und Sequenz-Konserviertheitsraten der zugehörigen TFBSs. Jeder Kreis steht für eine PSSM. Es besteht keine signifikante Korrelation (Spearman'scher Korrelationskoeffizient $\rho = 0.092$, $p = 0.417$) zwischen der Sequenz-Konserviertheitsrate der TFBSs einer PSSM und dem durchschnittlichen Informationsgehalt der PSSM.

mit niedriger Konserviertheitsrate (Sp1, GATA-1) wurde experimentell belegt, dass sie in Clustern auftreten (Hardison et al., 1993; Hermfisse et al., 1996). Weitere Beispiele sind der TRANSFAC®-Datenbank zu entnehmen: Für das *MT2A*-Gen des Menschen sind vier AP-2αA TFBSs im Bereich zwischen -230 und -110 bp stromaufwärts der TSS annotiert, und das *SFRS2*-Gen des Menschen besitzt elf c-Myb TFBSs im Bereich zwischen -940 und +140 bp relativ zur TSS. Abbildung 4.12 zeigt exemplarisch eine schematische Übersicht über die Konserviertheit dieser elf c-Myb TFBSs zwischen Mensch und Maus, Ratte, Kuh, Hund, Opossum sowie Gürteltier (die orthologen Sequenzen wurden für dieses Beispiel mittels der im Ensembl Release 38 enthaltenen Annotation bestimmt). Die Mehrzahl der TFBSs (sechs von elf) weist eine unterschiedliche Konserviertheit in den einzelnen Speziesvergleichen auf. Drei TFBSs sind in allen Speziesvergleichen konserviert, zwei TFBSs dahingegen in keinem Speziesvergleich. Diese TFBSs sind daher entweder spezifisch für den Menschen oder möglicherweise experimentelle Artefakte.

Eine weitere Erklärung für niedrige Konserviertheitsraten ist ein unterschiedliches Expressionsverhalten orthologer Gene in Mensch und Nagetier. Für die meisten Gene des β-Globin-Clusters ist beispielsweise bekannt (Ross Hardison, persönliche Mitteilung), dass

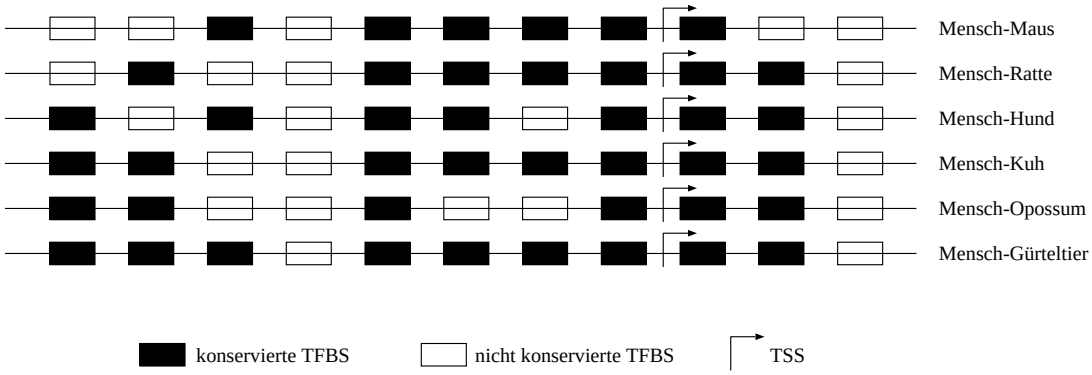

Abbildung 4.12: Schematische Darstellung der Konserviertheit von elf Mensch-TFBSs des TF c-Myb, die im Bereich von -940 bis +140 bp um die TSS des *SFRS2*-Gens liegen. Jede Zeile bezieht sich dabei auf ein paarweises Alignment, wobei konservierte bzw. nicht konservierte TFBSs durch gefüllte bzw. leere Kästchen repräsentiert werden. Sechs der elf TFBSs unterscheiden sich in ihrer Konserviertheit für die jeweiligen paarweisen Speziesvergleiche von Mensch mit Maus, Ratte, Kuh, Hund, Opossum oder Gürteltier. Für fünf TFBSs ist deren Konserviertheit zwischen Mensch und allen Spezies identisch.

sie in Mensch und Nagetier ein anderes Expressionsverhalten aufweisen und damit auch anderen Regulationsmechanismen unterliegen. Dies trifft für die Gene *HBB, Hbb-b1, HBD* und *HBG2* zu. Die geringe Konserviertheitsrate der TFBSs des TF GATA-1 kann darauf zurückgeführt werden, dass 23 der 50 untersuchten GATA-1 TFBSs stromaufwärts dieser Gene lokalisiert sind. Nur vier dieser 23 TFBS im β-Globin-Cluster sind konserviert ($C_{seq} = 17.4\%$). Von den verbleibenden 27 TFBS sind 18 konserviert und man erhält eine höhere Konserviertheitsrate von $C_{seq} = 66.7\%$, die sich signifikant von der Hintergrund-Konserviertheitsrate abhebt ($p = 9.2 \cdot 10^{-6}$). Es ist daher denkbar, dass die Konserviertheitsraten für einige TFs durch den Annotationsfokus von TRANSFAC® zu höheren oder niedrigeren Werten verzerrt sind.

4.2.2 Abhängigkeit der Konserviertheitsrate von der Genfunktion

Einige Studien (Bejerano et al., 2004; Choi et al., 2005; Lee et al., 2005; Sandelin et al., 2004; Woolfe et al., 2005) stellten eine Häufung von hoch konservierten Sequenzen in der Nähe von Genen, die in der Regulation von Transkription und Entwicklung eine Rolle spielen, fest. Bejerano et al. (2004) entdeckten 254 zu 100% konservierte nicht-codierende Sequenzen, die dieses Verhalten aufweisen. Diese Sequenzen sind länger als 200 bp und stimmen zu 100% zwischen den Genomen von Mensch, Maus und Ratte überein. Die meisten dieser Sequenzen sind auch in den Genomen von Huhn und Hund und ungefähr zwei Drittel sogar im Pufferfisch (*Fugu rubripes*) konserviert. Sandelin et al. (2004) führten eine

vergleichbare Studie mit weniger restriktiven Einstellungen zur Bestimmung der Konserviertheit durch und entdeckten 3583 ultra-konservierte Regionen ("ultra-conserved regions", UCRs). Eine UCR ist dabei als eine nicht-codierende, genomische Region definiert, die einen PI-Wert von über 95% in einem „sliding window"-Vergleich (Länge von 50 bp) zwischen Mensch und Maus besitzt und gleichzeitig mit Sequenzen überlappt, die zwischen den Genomen von Mensch und Pufferfisch konserviert sind. Das Auftreten dieser UCRs korreliert stark mit Genen, die für Schlüsselregulatoren der Vertebraten-Entwicklung, insbesondere TF-Gene, codieren. In einer anderen Studie identifizierten Woolfe et al. (2005) mittels eines Genom-Vergleiches von Mensch und Pufferfisch 1373 hoch konservierte, nicht-codierende Sequenzen. Die meisten dieser Sequenzen liegen in der Nähe von Genen, die die Entwicklung regulieren. Aufgrund der hohen evolutionären Divergenz zwischen Mensch und Pufferfisch ist anzunehmen, dass diese Sequenzen auch in allen anderen Vertebraten-Spezies existieren und essentiell für die Entwicklung von Vertebraten sind. Diese Hypothese wird von der Tatsache unterstützt, dass die Autoren keine ähnlichen Sequenzen in Genomen von Invertebraten fanden.

Die in den obigen Studien beschriebenen hoch konservierten Sequenzen liegen in den meisten Fällen mehrere kb von bekannten Genen entfernt und agieren womöglich als Enhancer oder spielen eine Rolle in der Strukturierung der genomischen Architektur. Im Gegensatz dazu konzentrierten sich Iwama und Gojobori (2004) in ihrer Studie auf die Konserviertheit der Sequenzen direkt stromaufwärts von 3055 orthologen Gen-Paaren zwischen Mensch und Maus. Sie fanden heraus, dass TF-Gene und Gene, die eine Rolle in der Entwicklung spielen, eine sehr hohe Konserviertheit ihrer Promotoren aufweisen, während Gene, die in Stoffwechsel und Zellzyklus involviert sind, eine geringe Konserviertheit ihrer Promotoren zeigen.

Während Iwama und Gojobori (2004) die Konserviertheit von Promotoren im Allgemeinen untersuchten, wurde in dieser Arbeit explizit überprüft, ob es eine Korrelation zwischen der Gen-Funktion und der Konserviertheit von einzelnen TFBSs gibt. Dazu wurden die TFBSs aus Datensatz I entsprechend der Funktion ihrer zugehörigen Gene, die anhand von „GO slim"-Begriffen beschrieben wird, gruppiert und die jeweilige Konserviertheitsrate bestimmt (siehe 3.3.7, S. 44). Eine TFBS konnte dabei mehreren Gen-Funktionen zugeordnet sein. Tabelle 4.2 gibt eine Übersicht über die Ergebnisse dieser Untersuchung.

TFBSs von Genen, die die Transkription regulieren ($C_{seq} = 80.5\%$), die beim Zelltod (77.9%), der Regulation biologischer Prozesse (77.3%), dem Binden an Nukleinsäuren (77.2%), der Signal-Weiterleitung (75.8%) und in der Entwicklung (75.1%) eine Rolle spielen, weisen hohe Konserviertheitsraten auf. TFBSs von Genen, die am Transport (62.1%) beteiligt sind oder eine gewisse katalytische Aktivität besitzen, wie z.B. Kinase- (66.9%),

Transferase- (61.9%) und Oxidoreduktase-Aktivität (59.9%), weisen dagegen niedrige Konserviertheitsraten auf.

Tabelle 4.2: Sequenz-Konserviertheitsrate für TFBSs von Genen bestimmter Funktion (für AVID-Alignments der OSPs aus Datensatz I)

„GO slim"-Begriff	N^*	C_{seq}	$p_{seq}{}^\dagger$
transcription regulator activity	262	80.5%	3.1E-49
hydrolase activity	199	79.4%	2.1E-35
cell death	195	77.9%	2.4E-32
regulation of biological process	745	77.3%	5E-120
nucleic acid binding	356	77.2%	3.8E-57
catabolism	187	77.0%	1.4E-29
signal transducer activity	714	75.8%	2E-106
development	526	75.1%	9.5E-76
protein binding	779	74.8%	1E-110
response to stimulus	703	74.8%	8E-100
cell communication	602	74.1%	3.1E-82
enzyme regulator activity	112	72.3%	9.1E-14
metabolism	1197	71.1%	9E-139
catalytic activity	646	67.5%	1.5E-60
kinase activity	136	66.9%	3.6E-12
cell motility	64	65.6%	4.5E-05
electron transport	95	65.3%	1.6E-07
transport	388	62.1%	2.8E-25
transferase activity	223	61.9%	3.2E-14
oxidoreductase activity	187	59.9%	3.6E-10
transporter activity	350	59.7%	6.2E-19
electron transporter activity	62	56.5%	0.03223

Diese Ergebnisse stehen in gutem Einklang mit den oben erwähnten Studien und deuten darauf hin, dass TFBSs, die Gene regulieren, die essentiell für die Entwicklung von Wirbeltieren sind, unter einem höheren evolutionären Druck stehen und daher stärker konserviert sind.

4.2.3 Positionsspezifität der Sequenz-Konserviertheit

Die einzelnen Positionen innerhalb von TFBSs sind von unterschiedlicher Bedeutung für die Ausbildung einer DNA-Protein-Bindung. Mirny und Gelfand (2002) untersuchten die

*Anzahl der TFBSs, die mit dem „GO slim"-Begriff verknüpft sind.

†Die Wahrscheinlicheit p, die beobachtete (oder eine größere) Differenz zwischen C_{seq} und C_{seq}^{bg} zufällig zu erhalten, wurde mit Gleichung (3.7) berechnet. Die p-Werte wurden Bonferroni-korrigiert.

Strukturen von Protein-DNA-Komplexen und die Sequenzen experimentell bestimmter TFBSs in *Escherichia coli*. Für die einzelnen Basenpaare in den TFBSs wurde der Informationsgehalt mit Gleichung (3.2) (siehe 3.2.2, S. 29) berechnet, der die Bedeutung jeder Position in der spezifischen Erkennung der TFBS widerspiegelt. Das Hauptergebnis dieser Studie war, dass der Informationsgehalt der einzelnen Positionen innerhalb von TFBSs mit der Anzahl von Protein-Kontakten signifikant positiv korreliert ist. D.h. innerhalb einer Spezies sind die Basenpaare, die mehr Kontakte mit dem TF aufweisen, evolutionär stärker konserviert.

Moses et al. (2003) machten eine ähnliche Beobachtung für TFBSs in *Saccharomyces cerevisiae* beim Vergleich mit drei anderen Hefe-Spezies. In dieser Studie wurde gezeigt, dass die Positionen innerhalb von TFBSs unterschiedliche Evolutionsraten aufweisen. Diese Variationen spiegeln die unterschiedliche Bedeutung einzelner Positionen für die Etablierung der TF-DNA-Bindung wider. Protein-DNA-Kristallstrukturen belegen, dass die evolutionär konservierten Positionen Kontaktpunkte für die Ausbildung der spezifischen Protein-DNA-Bindung sind. Moses et al. (2003) zeigten auch, dass die positionspezifische Evolutionsrate negativ mit dem Informationsgehalt von TFBSs korreliert ist, d.h. die variableren Positionen in der PSSM, deren Beitrag zur Stabilität der Protein-DNA-Bindung gering ist, weisen im Spezies-Vergleich mehr Mutationen auf.

Im Rahmen dieser Arbeit wurde überprüft, ob eine ähnliche Beobachtung für Mensch-Maus/Ratte-Vergleiche gemacht werden kann. Für 42 PSSMs, die in Datensatz I durch mindestens 15 TFBSs repräsentiert sind, wurde für alle Positionen der Informationsgehalt und die Konserviertheitsrate für AVID-Alignments berechnet. Die Konserviertheitsrate pro Position wurde mit den Gleichungen 3.5 und 3.6 berechnet (siehe 3.3.5, S. 43). Da jeweils nur eine Position der TFBS untersucht wurde, wurde die Länge $l = 1$ und der Konserviertheit-Schwellenwert $\mathrm{KS} = 100\%$ gesetzt. Die Abbildung 4.13 zeigt exemplarisch den Informationsgehalt und die Konserviertheitsrate pro Position für die TRANSFAC®-PSSMs mit den Zugriffsnummern M00925 (AP-1) bzw. M00932 (Sp1). Die Positionen mit einem höheren Informationsgehalt sind generell evolutionär stärker konserviert. Dies ist ein Hinweis darauf, dass diese Positionen entscheidend für die spezifische Ausbildung der Protein-DNA-Bindung sind. Durch ihre Bedeutung für die Stabilität des Protein-DNA-Komplexes sind sie einem höheren evolutionären Druck ausgesetzt und weisen weniger Mutationen auf. Es ist jedoch auffällig, dass einige Positionen trotz geringem Informationsgehalt (z.B. die Positionen 5 bis 8 der PSSM mit der Zugriffsnummer M00925, siehe Abbildung 4.13) hoch konserviert sind.

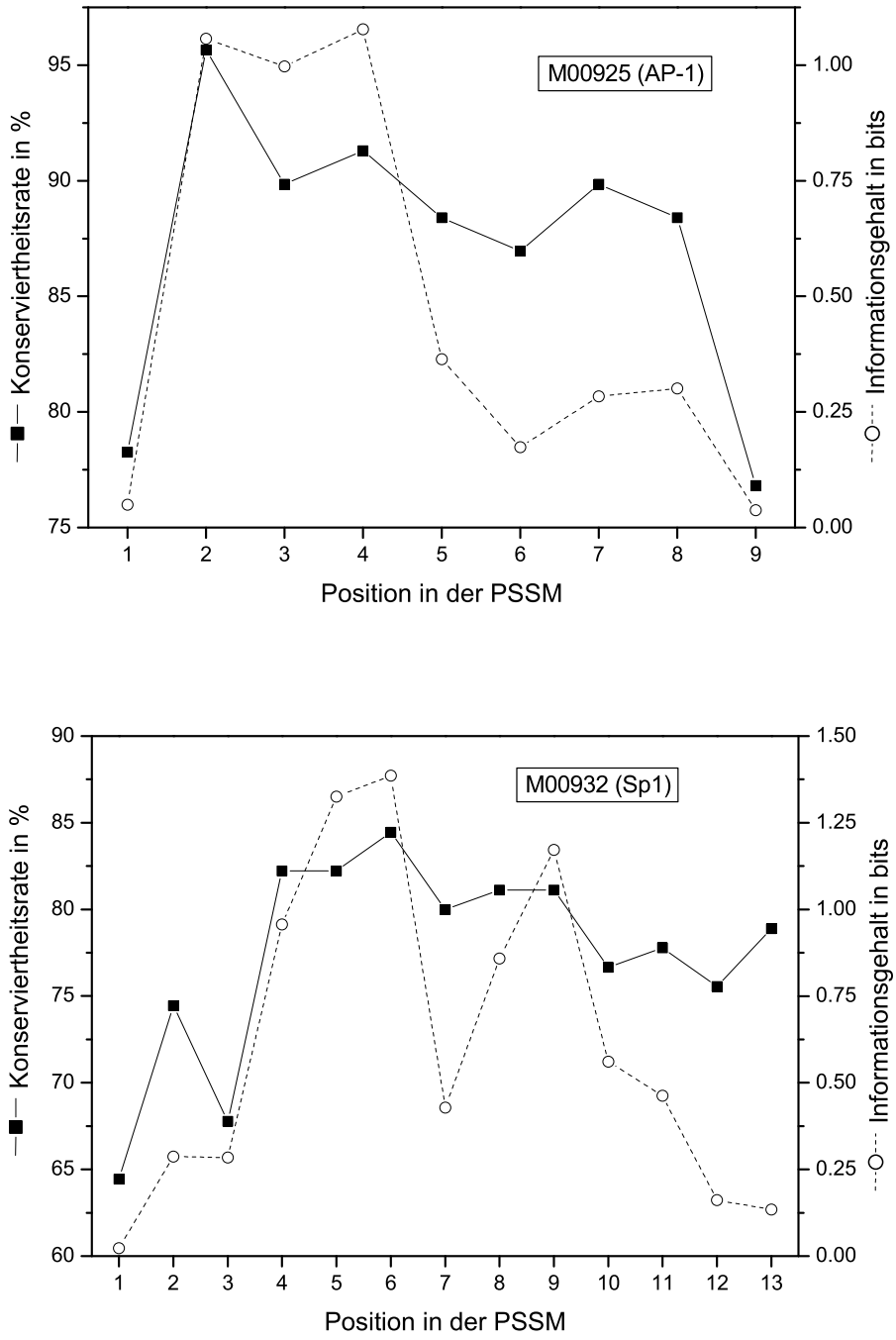

Abbildung 4.13: Positionsabhängigkeit von Informationsgehalt (offene Kreise) und Konserviertheitsrate (schwarze Quadrate) einzelner Basenpaare für die PSSMs mit den Zugriffsnummern M00925 (AP-1) und M00932 (Sp1). Die beiden Größen sind positiv korreliert, d.h. Positionen mit höherem Informationsgehalt sind häufig auch stärker konserviert.

Im Prinzip ist an diesen Positionen eine variable Sequenz-Erkennung durch den TF möglich, aber im jeweiligen Kontext scheint die bestimmte Sequenz für die Feinregulation der Genexpression von Bedeutung zu sein. Es ist denkbar, daß nur minimale Änderungen der Bindungsaffinität des TF durch Substitutionen an Positionen mit geringem Informationsgehalt zu Abweichungen von der idealen Expressionsstärke eines regulierten Gens führen könnten, weshalb auch diese Positionen evolutionär erhalten wurden.

Für jede der 42 PSSMs wurde überprüft, ob es eine Korrelation zwischen dem Informationsgehalt und der Konserviertheitsrate pro Position gibt. Dazu wurde der Spearman'sche Korrelationskoeffizient ρ (Press et al., 1992) mit dazugehörigen p-Werten berechnet (Implementation in R, http://www.r-project.org). Für nahezu alle PSSMs erhält man positive Werte für ρ (siehe Tabelle 4.3), wobei für 13 PSSMs der p-Wert kleiner gleich 0.05 ist. Insgesamt gibt es nur vier PSSMs mit einem negativen Wert für ρ, wobei keiner dieser Werte statistisch signifikant ist. Beispielsweise weist die PSSM mit der TRANSFAC®-Zugriffsnummer M00963 (T3R) mit $\rho = -0.4518$ den kleinsten Wert für den Spearman'schen Korrelationskoeffizienten auf, der zugeordnete p-Wert beträgt jedoch 0.2298. Dies bedeutet, dass zwar nur für wenige PSSMs statistisch relevante Korrelationen zwischen dem Informationsgehalt und der Konserviertheitsrate pro Position gefunden werden konnten, diese Korrelationen aber stets positiv sind. Dies ist ein Hinweis dafür, dass trotz einer generellen Konserviertheit von TFBSs die Positionen, die entscheidend für die Spezifität der Protein-DNA-Wechselwirkung sind, einem höheren evolutionären Druck unterliegen.

Tabelle 4.3: Spearman-Korrelation zwischen relativer Häufigkeit konservierter Basenpaare und Informationsgehalt für TRANSFAC®-PSSMs

ID*	TF	$N^{\dagger}$	$\lambda^{\ddagger}$	$\rho^{\S}$	p-Wert¶
M00925	AP-1	58	9	0.908	0.0013
M00932	Sp1	77	13	0.854	0.0003
M01007	SRF	26	19	0.806	3.47E-005
M00935	NF-AT	18	10	0.789	0.0098
M00971	Ets	27	8	0.764	0.0368
M00922	SRF	19	15	0.760	0.0015
M00926	AP-1	54	8	0.748	0.0368
M00976	AHR/HIF	27	9	0.733	0.0311
M00797	HIF-1	17	14	0.733	0.0038
M00801	CREB	18	6	0.705	0.1361
M00810	SRF	19	18	0.685	0.0022
M00933	Sp1	57	10	0.675	0.0372
M00761	p53	28	10	0.651	0.0465
M01032	HNF4	28	6	0.638	0.1750
M00806	NF-1	19	17	0.561	0.0207
M00924	AP-1	33	12	0.531	0.0771
M00774	NF-κB	15	16	0.479	0.0615
M01034	Ebox	55	10	0.471	0.1661
M01035	YY1	16	11	0.467	0.1456
M00775	NF-Y	19	13	0.467	0.1080
M00972	IRF	18	11	0.457	0.1543
M00789	GATA	49	7	0.449	0.3024
M01011	HNF1	19	21	0.425	0.0559
M00930	Oct-1	16	11	0.414	0.2031
M00770	C/EBP	32	12	0.327	0.2974
M00912	C/EBP	39	12	0.313	0.3195
M01031	HNF4	28	14	0.310	0.2770
M00982	KROX	15	14	0.271	0.3451
M00762	PPAR, HNF-4, COUP, RAR	16	13	0.256	0.3938
M00931	Sp1	114	10	0.220	0.5367
M00918	E2F	16	9	0.199	0.6134
M00799	Myc	15	7	0.139	0.7825
M00919	E2F	20	11	0.119	0.7240
M00981	CREB/ATF	24	9	0.117	0.7756
M00920	E2F	24	12	0.110	0.7328
M00795	Octamer	16	11	0.092	0.7861
M00915	AP-2	16	13	0.027	0.9278
M00790	HNF-1	17	18	0.009	0.9705
M00939	E2F-1	20	9	-0.038	0.9116
M01033	HNF4	28	6	-0.058	0.9194
M01029	TFE	20	8	-0.236	0.5821
M00963	T3R	19	9	-0.452	0.2298

*Zugriffsnummer der PSSM in TRANSFAC®.

†Anzahl der zur PSSM gehörenden TFBSs in Datensatz I.

‡Länge der PSSM in bp.

§Spearman'scher Korrelationskoeffizient.

¶Wahrscheinlichkeit, den Spearman'schen Korrelationskoeffizienten zufällig zu erhalten.

4.3 Muster-Konserviertheit von Transkriptionsfaktor-Bindestellen

Die Sequenz-Konserviertheit von TFBSs unterscheidet sich für verschiedene TFs, wie in Abschnitt 4.2.1 (S. 79) gezeigt wurde. Da die DNA-Bindungs-Spezifität von TFs recht gering ist, kann die Sequenz einer TFBS in einem Alignment zwar nur schwach konserviert sein, der TF aber dennoch an der orthologen Sequenz binden. Andererseits kann eine TFBS zwar hoch Sequenz-konserviert sein, aber eine einzige Subsitution, Insertion oder Deletion trotzdem die Funktionalität der orthologen TFBS zerstört haben. Deshalb wurde im Rahmen dieser Arbeit zusätzlich zur Sequenz-Konserviertheit auch die Konserviertheit des Musters einer TFBS überprüft.

Für 1990 der 2678 untersuchten TFBSs aus Datensatz I existiert in TRANSFAC® mindestens eine korrespondierende PSSM. Für diese TFBSs wurde die Muster-Konserviertheit bestimmt (siehe 3.3.8, S. 45). Dazu wurden alle OSPs nach Treffern korrespondierender PSSMs durchsucht, die nahezu an derselben Position im Alignment liegen und einen ähnlichen MSS haben. Fielen solche Treffer auf eine bekannte TFBS, wurde diese als Muster-konserviert bezeichnet.

93 der 100 Nagetier-TFBSs, die zu Mensch-TFBSs des gleichen TF aligniert wurden, sind Muster-konserviert und wurden nicht in die Berechnung der Muster-Konserviertheitsrate einbezogen, um keine Informationen doppelt zu zählen. Die Muster-Konserviertheitsrate C_{pat} für die verbleibenden 1890 TFBSs beträgt 69.5%. Von diesen 1890 TFBSs sind 72.3% auf Sequenzebene (für einen KS von 65%) konserviert. 58.4% sind auf Muster- und Sequenzebene konserviert, d.h. 13.9% sind nur auf Sequenz- und 11.1% nur auf Musterebene konserviert. 16.6% sind weder auf Muster- noch auf Sequenzebene konserviert. Tabelle A.1 (siehe Anhang, S. 123) zeigt eine Übersicht der Ergebnisse für Datensatz I.

4.3.1 Abhängigkeit der Konserviertheitsrate vom Transkriptionsfaktor

Die Muster-Konserviertheitsrate wurde wie in Abschnitt 4.2.1 (S. 79) für die TFBSs jedes TF einzeln bestimmt. Für die TFBSs der meisten TFs wurden ähnliche Sequenz- und Muster-Konserviertheitsraten erhalten, für einige TFs unterscheiden sie sich jedoch recht deutlich (siehe Tabelle 4.4).

Tabelle 4.4: Muster-Konserviertheitsrate für TFBSs bestimmter TFs (für AVID-Alignments der OSPs aus Datensatz I)

TF	N^*	$C_{seq}{}^\dagger$	$p_{seq}{}^\ddagger$	C_{pat}	C_{pat}^{bg}	$p_{pat}{}^\ddagger$
MyoD	16	100.0%	5.3E-06	100.0%	21.21%	1.5E-09
MEF-2A	13	100.0%	0.00012	92.3%	19.91%	3.8E-06
SRF	33	96.9%	6.4E-12	87.8%	24.17%	1.7E-12
NF-AT1	31	96.7%	4.9E-11	93.5%	28.31%	2.9E-12
USF-1	22	86.3%	0.0001	77.2%	19.70%	8.8E-07
NF-κB	38	84.2%	6.7E-08	71.0%	21.98%	1.4E-08
CREB	49	83.6%	3.8E-10	81.6%	28.62%	1.8E-12
c-Ets-1	18	83.3%	0.00368	66.6%	28.49%	0.07935
ER-α	17	82.3%	0.00876	52.9%	22.08%	0.48353
MITF	24	79.1%	0.00128	54.1%	18.42%	0.00828
C/EBPβ	42	78.5%	1.1E-06	90.4%	33.40%	1.7E-12
Sp3	27	77.7%	0.00073	92.5%	30.26%	1.7E-09
RXR-α	54	75.9%	1.2E-07	64.8%	27.12%	7.5E-07
HNF-1α	20	75.0%	0.03155	85.0%	30.95%	8.2E-05
AP-1	77	74.0%	4.2E-10	76.6%	28.27%	1.8E-16
YY1	29	72.4%	0.00469	68.9%	28.87%	0.00083
POU2F1	32	65.6%	0.04194	87.5%	29.05%	8.3E-10
Sp1	229	63.7%	1.5E-16	78.1%	31.94%	8.6E-45
C/EBPα	87	63.2%	9.3E-06	89.6%	39.27%	1.2E-20
c-Myb	16	62.5%	1	25.0%	28.19%	1
p53	27	51.8%	1	55.5%	17.85%	0.00107
Nkx2-1	33	51.5%	1	66.6%	33.35%	0.00862
AP-2αA	29	48.2%	1	89.6%	33.59%	5E-08
GATA-1	50	44.0%	1	50.0%	21.71%	0.0009

Es gibt zwei grundlegende Erklärungen, warum eine TFBS Sequenz-, aber nicht Muster-konserviert bzw. Muster-, aber nicht Sequenz-konserviert ist:

1. Da die Konserviertheit auf Sequenzebene nicht zwangsläufig einen PI-Wert von 100% verlangt, ist es denkbar, dass in einer ansonsten hoch konservierten TFBS nur einige wenige essentielle Nukleotide mutiert sind, wodurch allerdings eine Matrixsuche in der orthologen Sequenz keinen Treffer oberhalb des MSS-Schwellenwertes ergeben würde. Die TFBS wäre daher Sequenz-, aber nicht Muster-konserviert.

2. Anderenfalls ist es denkbar, dass eine sehr schwach Sequenz-konservierte TFBS in

*Anzahl der TFBSs pro TF.

†Die Hintergrund-Konserviertheitsrate C_{seq}^{bg} beträgt 35.22%.

‡Die Wahrscheinlicheit p, die beobachtete (oder eine größere) Differenz zwischen C_{seq} und C_{seq}^{bg} zufällig zu erhalten, wurde mit Gleichung (3.7) berechnet. Die p-Werte wurden Bonferroni-korrigiert.

der orthologen Sequenz trotzdem von einem TF gebunden wird, da der TF generell eine sehr geringe Bindungs-Spezifität aufweist. Die TFBS wäre in diesem Falle muster-, aber nicht Sequenz-konserviert.

Beispiele für den ersten Fall sind die TFBSs von ER-α, die auf Sequenzebene sehr gut konserviert sind ($C_{seq} = 82.3\%$), aber auf Musterebene nicht ($C_{pat} = 52.9\%$). Zum Beispiel zeigen die erhaltenen Alignments, dass in einer ER-α-TFBS im Calbindin D9K-Gen der Ratte ein essentielles Nukleotid der orthologen Sequenz in der TFBS mutiert ist, wodurch man einen kleinen Wert für den MSS erhält, obwohl die umgebende Sequenz konserviert ist (siehe Abbildung 4.14).

```
                mSS=0.997
Ratte       CAGGTCAGGGT
            ||||| || ||
Mensch      CAGGTTAGTGT
                mSS=0.842
```

Abbildung 4.14: Beispiel für eine Sequenz-, aber nicht Muster-konservierte TFBS. Die TFBS für den TF ER-α liegt stromaufwärts des Calbindin D9K-Gens der Ratte und ist an neun von elf Positionen konserviert.

Es ist allerdings möglich, dass die eingesetzte PSSM zu stringent ist, da sie aus nur 20 TFBSs gebildet wurde, und nicht die reale Variabilität für TFBSs dieses Faktors widerspiegelt. D.h. die Mutation eines Nukleotids in der TFBS muss den TF nicht zwangsweise daran hindern, *in vivo* zu binden. Als weitere Beispiele für TFs, deren TFBSs eine hohe Sequenz-, aber geringe Muster-Konserviertheit aufweisen, sind NF-κB und c-Ets-1 zu nennen (siehe Tabelle 4.4).

Beispiele für den zweiten Fall sind TFBSs für C/EBPα, AP-2αA und Sp1. C/EBPα zeigt eine geringe Sequenz-Konserviertheit ($C_{seq} = 63.2\%$), jedoch eine hohe Muster-Konserviertheit ($C_{pat} = 89.6\%$), die zumindest teilweise in der hohen Degeneriertheit von C/EBPα-TFBSs begründet liegen mag. Dies führt dazu, dass recht viele Sequenzabschnitte eine gewisse Ähnlichkeit zur korrespondierenden PSSMs aufweisen. Dies spiegelt sich auch in der hohen Hintergrund-Konserviertheitsrate ($C_{pat}^{bg} = 39.27\%$) für diesen TF wider. Auch AP-2αA und, in einem geringeren Ausmaß, Sp1 weisen eine hohe Muster- und eine geringe Sequenz-Konserviertheit auf (siehe Tabelle 4.4), wobei diese TFs geringere Hintergrund-Konserviertheitsraten als C/EBPα haben.

Die Konserviertheitsrate der TFBSs von AP-2αA unterscheidet sich auf Sequenzebene nicht signifikant von der Hintergrund-Konserviertheit (Bonferroni-korrigierten p-Wert von 1), wohingegen dieser Unterschied auf Musterebene mit einem Bonferroni-korrigierten p-

Wert von $5 \cdot 10^{-8}$ signifikant ist. Das Konzept der Muster-Konserviertheit ist daher für TFs nützlich, bei denen rein Sequenz-basiertes phylogenetisches Footprinting versagen könnte.

4.4 Cluster von Transkriptionsfakor-Bindestellen und deren Konserviertheit

Es ist bekannt, dass TFBSs häufig in Clustern auftreten (Frith et al., 2003; Hannenhalli und Levy, 2002), da TFs in Eukaryoten ihre regulatorische Wirkung erst durch ihre Kombination voll entfalten können. Um zu untersuchen, inwiefern TFBSs, die in Clustern vorkommen, konserviert sind, wurden TFBSs aus der TRANSCompel®-Datenbank untersucht (siehe 3.1.2, S. 26).

Die TRANSCompel®-Datenbank (siehe 3.1.2, S. 26) enthält Informationen über TFBSs, die in CEs vorkommen. CEs bestehen aus zwei benachbarten TFBSs, bei denen bekannt ist, dass die an sie bindenden TFs miteinander interagieren. Datensatz III enthält 405 solcher TFBSs. Diese TFBSs wurden analog zu den TFBSs aus Datensatz I auf ihre Konserviertheit in paarweisen Mensch-Maus/Ratte-Alignments untersucht. Die Alignments wurden mit T-COFFEE (siehe 3.2.3.10, S. 38) erstellt.

Betrachtet man die Konserviertheit der einzelnen TFBSs in einem CE, sind die PI-Werte der beiden TFBSs positiv korreliert. Abbildung 4.15 zeigt die Häufigkeit von PI-Wertepaaren in den untersuchten CEs. In den meisten Fällen sind beide TFBSs sehr stark konserviert, für 53 CEs sind sogar beide TFBSs zu 100% konserviert.

Die Verteilung der PI-Werte der TFBSs aus Datensatz III in der TRANSCompel®-Datenbank unterscheidet sich von der der TFBSs aus Datensatz I in der TRANSFAC®-Datenbank (siehe Abbildung 4.16) dadurch, dass der Anteil moderat konservierter TFBSs für die TRANSFAC®-Datenbank höher ist, während die TRANSCompel®-Datenbank wesentlich mehr TFBSs enthält, die zu 100% Sequenz-konserviert sind (42% gegenüber 26%).

Eine Erklärung für die Diskrepanz in der Verteilung der PI-Werte zwischen den Datensätzen I und III könnte eine unterschiedliche Qualität der Annotation in den Datenbanken sein. Es wäre möglich, dass die TFBSs in der TRANSCompel®-Datenbank verlässlicher annotiert sind, da für sie zusätzlich zur TF-DNA-Wechselwirkung eine TF-TF-Wechselwirkung von funktioneller Bedeutung (d.h. Aktivierung oder Repression der Transkription) nachgewiesen sein muss, damit sie in die Datenbank aufgenommen werden. Die TFBSs in der TRANSFAC®-Datenbank sind mit einem Qualitätswert versehen (siehe Tabelle 3.1, S. 26), der durch die experimentelle Methode ihrer Bestimmung vorgegeben ist. Für die TFBSs jedes Qualitätswerts wurde eine eigene PI-Wert-Verteilung berechnet (Qualität 1: 582 TFBSs, 2: 582, 3: 387, 4: 235, 5: 109, 6: 333; für 443 TFBSs war kein Qualitätswert annotiert, diesen wurde der Qualitätswert „NULL“ zugeordnet). Für keinen Qualitätswert ist eine Verteilung zu beobachten, die der der TFBSs in der TRANSCompel®-Datenbank nahe kommt (Abbildung 4.17). Selbst von den 331 TFBSs

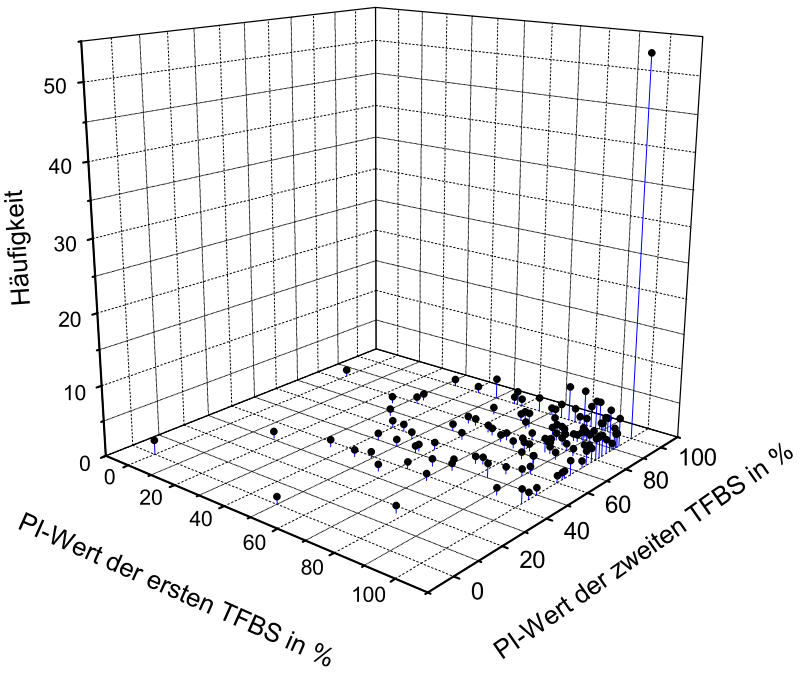

Abbildung 4.15: „Stecknadelplot“ der Häufigkeiten von PI-Wertepaaren in CE. Auf der x- und y-Achse sind die PI-Werte der einzelnen TFBSs eines CEs aufgetragen, die z-Achse gibt die Häufigkeit dieses PI-Wertepaares an. Der Abbildung ist zu entnehmen, dass bevorzugt beide TFBSs in einem CE stark konserviert sind.

mit Qualtitätswert 1, der für funktionell bestätigte TFBSs und damit die verlässlichste TRANSFAC®-Annotation steht, sind nur ca. 31% der TFBSs zu 100% Sequenz-konserviert.

Die Unterschiede in den PI-Wert-Verteilungen zwischen den Datenbanken sind daher wahrscheinlich nicht auf Unterschiede in der Qualität der Annotation zurückzuführen, sondern deuten darauf hin, dass durch die funktionelle Interaktion von TFs ein höherer evolutionärer Druck auf die zugehörigen TFBSs ausgeübt wird. Dies spiegelt sich in hohen PI-Werten dieser TFBSs wider.

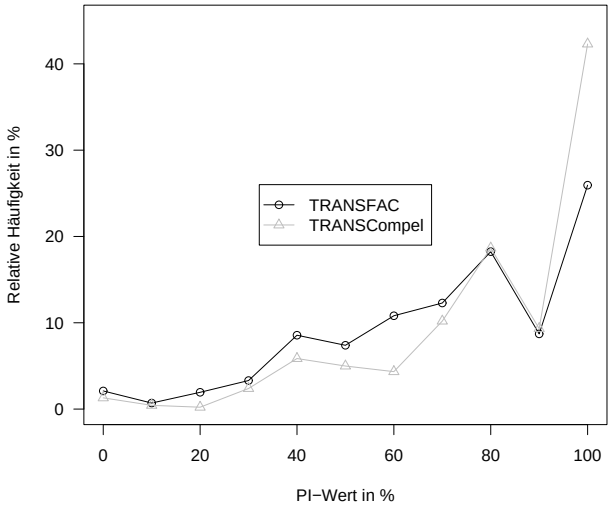

Abbildung 4.16: Die PI-Wert-Verteilungen der TFBSs in TRANSFAC® und TRANSCompel® unterscheiden sich dadurch, dass in TRANSCompel® die meisten TFBSs hoch konserviert sind, während in TRANSFAC® der Anteil schwächer konservierter TFBSs größer ist. Besonders auffällig ist der Unterschied bei den TFBSs, die zu 100% konserviert sind. Dies sind ca. 42% der TRANSCompel®- und ca. 26% der TRANSFAC®-TFBSs.

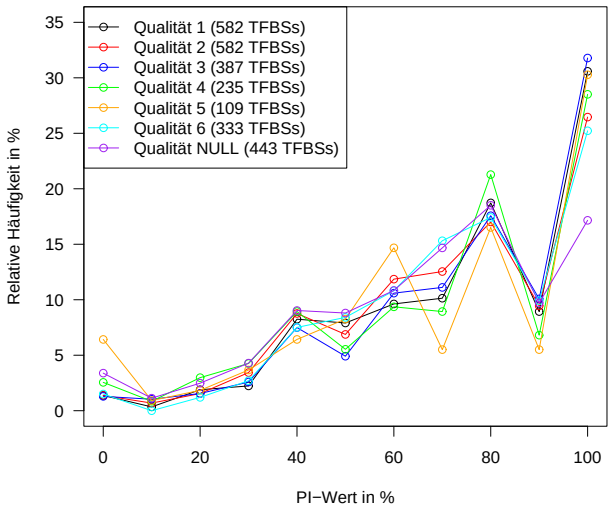

Abbildung 4.17: PI-Wert-Verteilungen der TFBSs in TRANSFAC®, aufgeschlüsselt nach ihrem Qualitätswert. Die Verteilung für den Qualitätswert „NULL“ weist deutlich weniger zu 100% Sequenz-konservierte TFBSs auf als die anderen Verteilungen. Für die Qualitätswerte von 1 bis 5 sind im Mittel 29% der TFBSs zu 100% konserviert.

4.5 Verschiedene Spezies im paarweisen Alignment

Die vorigen Untersuchungen konzentrierten sich auf Sequenz-Vergleiche zwischen Mensch und Maus bzw. Ratte. Die Sequenzen dieser Genome sind schon seit längerem verfügbar und wurden daher häufig in der vergleichenden Genomik eingesetzt. In den letzten Jahren ist die Anzahl der verfügbaren Genome gestiegen, sodass die Nützlichkeit anderer Spezies zur Detektion von *cis*-regulatorischen Elementen im menschlichen Genom evaluiert werden kann.

Im Rahmen dieser Arbeit wurde untersucht, inwiefern paarweise Vergleiche zwischen Mensch und Kuh, Hund, Maus bzw. Ratte geeignet sind, um experimentell bekannte TFBSs des Menschen zu identifizieren. Der Fokus wurde auf das menschliche Genom als Bezugspunkt gelegt, da der Anteil an annotierten TFBSs in TRANSFAC® für den Menschen am größten ist. Für Datensatz IV, der aus 928 Mensch-TFBSs in 234 OSPs besteht, wurden orthologe Sequenzen in Kuh, Hund, Maus und Ratte gefunden (siehe 3.3.3, S. 41). Diese OSPs wurden für jede Spezies mit DIALIGN (siehe 3.2.3.7, S. 36) aligniert, und die Konserviertheitsrate C_{seq} sowie die Hintergrund-Konserviertheitsrate C_{seq}^{bg} in Abhängigkeit vom KS bestimmt.

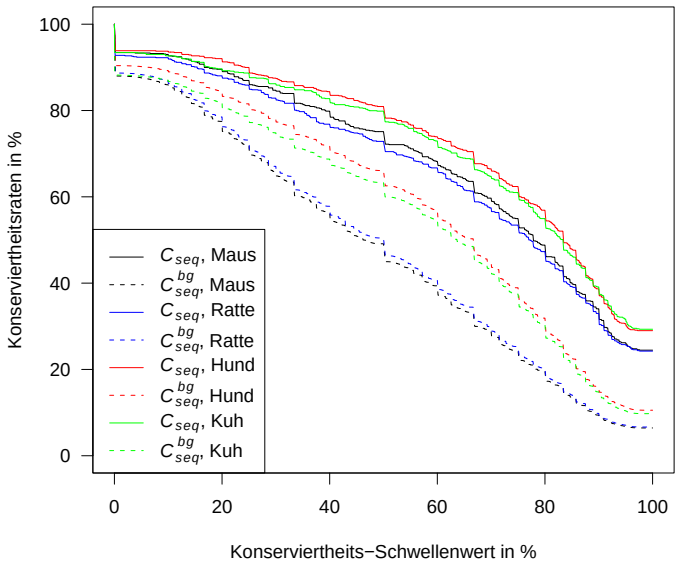

Abbildung 4.18: Die Konserviertheitsraten in paarweisen Alignments mit Mensch-Sequenzen (Datensatz IV, 928 TFBSs) sind für Vergleiche mit Hund oder Kuh generell höher als für Vergleiche mit Maus oder Ratte. Der Unterschied zwischen C_{seq} und C_{seq}^{bg} ist für Hund und Kuh kleiner als für Maus und Ratte.

Für die Alignments mit Kuh und Hund wurden für sowohl C_{seq} als auch für C_{seq}^{bg} (siehe Abbildung 4.18) aufgrund des im Vergleich zu Maus und Ratte geringeren evolutionären Abstandes zum Menschen generell höhere Werte erhalten. Ein phylogenetischer Baum, der die Verwandschaftsverhältnisse dieser Spezies beschreibt, findet sich z.B. in Thomas et al. (2003). Nicht-funktionelle Sequenzabschnitte der Spezies Kuh und Hund haben daher im Vergleich zum Menschen weniger Mutationen als Maus oder Ratte angesammelt. Eine Auftragung in einem ROC-Diagramm (siehe Abbildung 4.19) zeigt, dass Mensch-Maus-Alignments die beste Diskrimination zwischen funktionellen TFBSs und Hintergrund-Sequenzen erzielen, danach folgen Mensch-Ratte-Alignments und gleichauf Mensch-Hund- und Mensch-Kuh-Alignments.

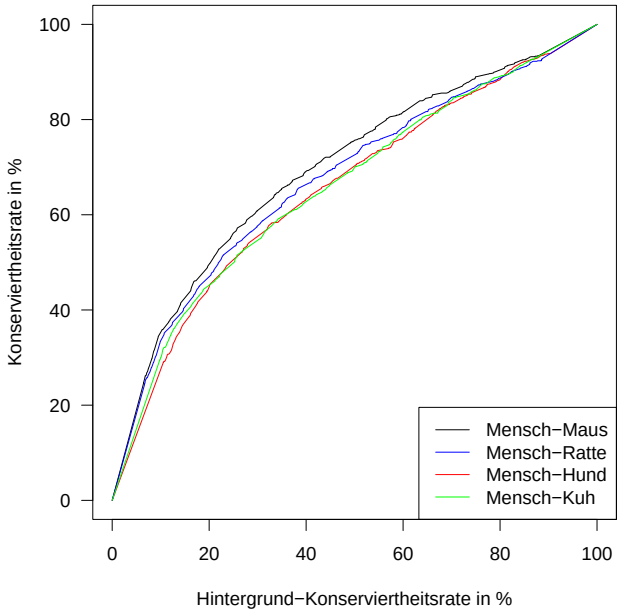

Abbildung 4.19: ROC-Diagramm für paarweise DIALIGN-Alignments zwischen orthologen Mensch- und Maus-, Ratte-, Hund- und Kuh-Sequenzen. Die untersuchten 928 Mensch-TFBSs heben sich in ihrer Konserviertheit von den Hintergrund-Sequenzen ab. Dieser Unterschied ist für Mensch-Maus-Vergleiche am höchsten.

Optimale KSs

Die optimalen KSs unterscheiden sich für die einzelnen Spezies, wie man Abbildung 4.20 entnehmen kann. Die hohen Werte von 76% bzw. 78% für Kuh bzw. Hund belegen noch einmal, dass die allgemeine Konserviertheit zwischen Mensch und diesen beiden Spezies sehr hoch ist. Für Maus bzw. Ratte liegen die optimalen KSs bei 65% bzw. 64% (siehe dazu

auch 4.2, S. 73). Mit steigendem evolutionären Abstand der mit dem Menschen verglichenen Spezies ist zu erwarten, dass der optimale KS sinkt und die Diskrimination zwischen TFBSs und Hintergrund-Sequenzen sich bis zu einem optimalen evolutionären Abstand verbessert und danach aufgrund schlechter werdender Alignmentqualität wieder geringer wird.

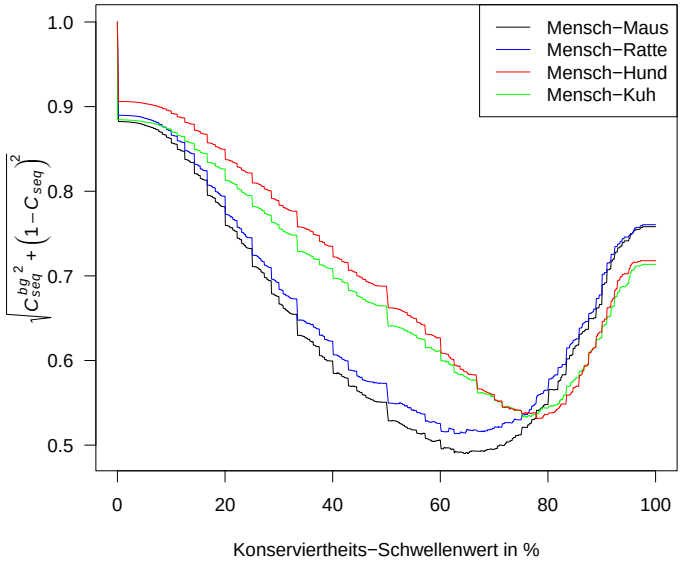

Abbildung 4.20: Für jeden KS und jede Spezies wurde der Abstand zur linken oberen Ecke des ROC-Diagramms berechnet. Je kleiner der Abstand, um so besser ist der Kompromiss zwischen der Konserviertheitsrate und falsch positiven Vorhersagen. Die optimalen Schwellenwerte sind für Maus (65%) und Ratte (64%) niedriger als für Hund (78%) und Kuh (76%).

Spezies-Spezifität der Konserviertheit

Um zu untersuchen, ob eine Spezies-Spezifität der Konserviertheit der 928 Mensch-TFBSs aus Datensatz IV existiert, wurden die PI-Werte aller TFBSs aus den verschiedenen Spezies-Vergleichen gegeneinander aufgetragen (siehe Abbildung B.2, S. 130). Insgesamt ist dabei eine positive Korrelation zu erkennen, die für Vergleiche zwischen Mensch und Maus bzw. Ratte besonders hoch (Pearson Korrelationskoeffizient von $R = 0.74$) ist. Dies ist auf den geringen evolutionären Abstand zwischen Maus und Ratte zurückzuführen. Für die übrigen Speziesvergleiche erhält man Werte von R zwischen 0.38 und 0.44. Insgesamt erhält man für den Großteil der TFBSs ähnliche PI-Werte in den Speziesvergleichen: 90 der 928 TFBSs haben für keinen der vier Speziesvergleiche einen PI-Wert $\geq 65\%$ und 442 TFBSs

haben für alle vier Speziesvergleiche einen PI-Wert $\geq 65\%$.

Allerdings unterscheiden sich für einige TFBSs die PI-Werte in den einzelnen Speziesvergleichen: 111 TFBSs weisen in einem, 138 TFBSs in zwei und 147 TFBSs in drei der Speziesvergleiche einen PI-Wert $\geq 65\%$ auf. Würden solche TFBSs in eine bestimmte Gruppe fallen, wäre dies ein Hinweis auf eine Spezies-spezifische Konserviertheit. Um eine Abhängigkeit vom TF zu untersuchen, wurden für jeden Speziesvergleich die Konserviertheitsraten der TFBSs jedes TF, der durch mindestens zehn TFBSs in Datensatz IV vertreten ist, berechnet (siehe Tabelle 4.5, S. 99). Zur Berechnung wurden die erhaltenen optimalen KSs (siehe S. 97) verwendet.

Tabelle 4.5: Sequenz-Konserviertheitsraten von 928 Mensch-TFBSs (Datensatz IV) für Speziesvergleiche zwischen Mensch und Maus, Ratte, Kuh oder Hund, aufgeschlüsselt nach TF mit mindestens zehn TFBSs

TF	N^*	$C_{seq}^{\text{Maus}\dagger}$	$p_{seq}^{\text{Maus}\ddagger}$	$C_{seq}^{\text{Ratte}\dagger}$	$p_{seq}^{\text{Ratte}\ddagger}$	$C_{seq}^{\text{Kuh}\dagger}$	$p_{seq}^{\text{Kuh}\ddagger}$	$C_{seq}^{\text{Hund}\dagger}$	$p_{seq}^{\text{Hund}\ddagger}$
Sp1	74	67.6%	7.8E-08	71.6%	3.9E-09	54.1%	0.0112	51.4%	0.0218
AP-1	34	73.5%	6.7E-05	79.4%	3.9E-06	79.4%	2.7E-06	61.8%	0.0145
C/EBPα	33	42.4%	1	39.4%	1	36.4%	1	45.5%	1
c-Jun	26	80.8%	3.1E-05	80.8%	6.2E-05	80.8%	4.6E-05	73.1%	0.0009
AP-2αA	20	45.0%	1	40.0%	1	40.0%	1	35.0%	1
CREB	19	78.9%	0.0019	73.7%	0.0183	84.2%	0.0003	63.2%	0.1843
GATA-1	19	52.6%	1	42.1%	1	26.3%	1	31.6%	1
NF-κB	19	73.7%	0.0120	78.9%	0.0031	73.7%	0.0153	52.6%	1
c-Fos	18	72.2%	0.0266	72.2%	0.0392	77.8%	0.0059	66.7%	0.0967
NF-AT1	16	93.8%	2.5E-05	87.5%	0.0006	81.3%	0.0046	87.5%	0.0003
NF-κB1	14	57.1%	1	71.4%	0.1655	64.3%	0.6025	57.1%	1
c-Myb	14	50.0%	1	64.3%	0.6698	50.0%	1	50.0%	1
IRF-1	13	76.9%	0.0504	76.9%	0.0689	61.5%	1	84.6%	0.0052
HNF-4	12	66.7%	0.5598	58.3%	1	58.3%	1	66.7%	0.4823
HNF-4α	12	41.7%	1	41.7%	1	50.0%	1	66.7%	0.4823
MITF	12	100.0%	5.9E-05	100.0%	9.2E-05	100.0%	7.6E-05	75.0%	0.09732
POU1F1a	12	83.3%	0.0167	75.0%	0.1540	83.3%	0.0201	91.7%	0.0011
POU2F1	12	66.7%	0.5598	41.7%	1	41.7%	1	41.7%	1
RXR-α	12	41.7%	1	41.7%	1	41.7%	1	41.7%	1
RelA	12	66.7%	0.5598	75.0%	0.1540	75.0%	0.1367	50.0%	1
Sp3	12	75.0%	0.1167	83.3%	0.0232	75.0%	0.1367	41.7%	1
NF-IL6-2	11	63.6%	1	54.5%	1	63.6%	1	63.6%	1
C/EBPβ	10	60.0%	1	60.0%	1	50.0%	1	60.0%	1
Egr-1	10	70.0%	0.5834	90.0%	0.0148	80.0%	0.1186	50.0%	1
LEF-1	10	60.0%	1	60.0%	1	80.0%	0.1186	50.0%	1
NF-Y	10	70.0%	0.5835	60.0%	1	50.0%	1	60.0%	1
p53	10	70.0%	0.5835	50.0%	1	80.0%	0.1186	50.0%	1

*Anzahl der TFBSs pro TF.

†Die Hintergrund-Konserviertheitsrate C_{seq}^{bg} beträgt für Mensch-Maus-Vergleiche 33.67% (KS $= 65\%$), für Mensch-Ratte-Vergleiche 34.91% (KS $= 64\%$), für Mensch-Kuh-Vergleiche 34.37% (KS $= 76\%$) und für Mensch-Hund-Vergleiche 32.89% (KS $= 78\%$).

‡Die Wahrscheinlicheit p, die beobachtete (oder eine größere) Differenz zwischen C_{seq} und C_{seq}^{bg} zufällig

Dabei sind allerdings keine signifikanten Unterschiede festzustellen, die darauf hindeuten würden, daß ein bestimmter TF in Maus, Ratte, Kuh oder Hund seine Funktion verändert oder gar komplett eingebüßt hätte. Zwar existieren Beispiele, wo TFBSs eines TF sich in einem Speziesvergleich nicht signifikant von der Hintergrund-Konserviertheit unterscheiden, z.B. IRF-1 TFBSs im Mensch-Kuh-Vergleich ($C_{seq}^{\text{Hund}} = 61.5\%, p_{seq}^{\text{Hund}} = 1$), aber diese Unterschiede kommen dadurch zustande, daß die TFBSs bestimmter Gene unterschiedlich konserviert sind. So liegen fünf der 13 IRF-1 TFBSs stromaufwärts des *IFNB1*-Gens. Im Mensch-Kuh-Vergleich sind nur zwei dieser fünf TFBSs konserviert, in den anderen Speziesvergleichen hingegen alle fünf.

Da bekannt ist, daß sich die Regulation bestimmter Gene in einzelnen Spezies stark unterscheiden kann (siehe 4.2.1, S. 82), wurden in einer zweiten Analyse für jeden Speziesvergleich die Konserviertheitsraten der TFBSs jedes einzelnen Gens, für das mindestens fünf TFBSs im Datensatz IV enthalten sind, berechnet (siehe Tabellen 4.6 und A.2, S. 124). Einige Gene zeigen sehr unterschiedliche Konserviertheitsraten für die einzelnen Speziesvergleiche. Z.B. sind die 19 TFBSs des *IFNB1*-Gen im Mensch-Kuh-Vergleich nur schwach konserviert ($C_{seq}^{\text{Kuh}} = 57.9\%$), während sie in den übrigen Speziesvergleichen Konserviertheitsraten $\geq 89.5\%$ aufweisen. Weitere auffällige Beispiele sind die TFBSs der Gene *CSF2*, *CCND1*, *JUN* oder auch *TYR*. Abbildung 4.21 zeigt eine schematische Übersicht der Konserviertheit von zehn annotierten TFBSs des *JUN*-Gens für alle vier Speziesvergleiche. Im Mensch-Hund-Vergleich sind vier dieser TFBSs nicht konserviert.

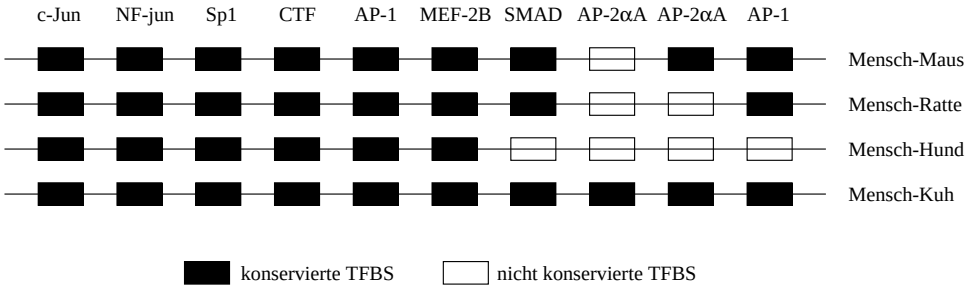

Abbildung 4.21: Schematische Übersicht der Konserviertheit von 10 TFBSs des *JUN*-Gens. Die Namen der zugehörigen TFs sind in der ersten Zeile angegeben. Es zeigen sich deutliche Unterschiede in der Konserviertheit der TFBSs in den einzelnen Speziesvergleichen: Im Mensch-Kuh-Vergleich sind alle 10 TFBSs konserviert, wohingegen im Mensch-Hund-Vergleich vier TFBSs nicht konserviert sind.

Eine genauere Untersuchung (siehe Abbildung B.1, S. 125) zeigte, dass sich für drei dieser TFBSs in der Hund-Sequenz Treffer für PSSMs des gleichen TF befinden, die im Alignment um 15 bis 50 Positionen verschoben sind. Dieses Beispiel belegt, dass sich die

zu erhalten, wurde mit Gleichung (3.7) berechnet. Die p-Werte wurden Bonferroni-korrigiert.

Positionierung bestimmter TFBSs, die für die Regulation eines Gens verantwortlich sind, im Laufe der Evolution Spezies-spezifisch verändert haben kann.

Tabelle 4.6: Sequenz-Konserviertheitsraten von 928 Mensch-TFBSs (Datensatz IV) für Speziesvergleiche zwischen Mensch und Maus, Ratte, Kuh oder Hund, aufgeschlüsselt nach Genen mit mindestens 5 TFBSs; Fortsetzung im Anhang (S. 124)

Gen	N*	$C_{seq}^{\text{Maus}\dagger}$	$p_{seq}^{\text{Maus}\ddagger}$	$C_{seq}^{\text{Ratte}\dagger}$	$p_{seq}^{\text{Ratte}\ddagger}$	$C_{seq}^{\text{Kuh}\dagger}$	$p_{seq}^{\text{Kuh}\ddagger}$	$C_{seq}^{\text{Hund}\dagger}$	$p_{seq}^{\text{Hund}\ddagger}$
HBG2	28	42.9%	1	35.7%	1	25.0%	1	25.0%	1
FOS	23	78.3%	0.0010	82.6%	0.0003	73.9%	0.0085	78.3%	0.0007
ADH1B	22	22.7%	1	9.1%	1	18.2%	1	13.6%	1
IFNB1	19	94.7%	2.7E-06	89.5%	8.8E-05	57.9%	1	94.7%	1.8E-06
IL2	17	100.0%	6.2E-07	94.1%	3.7E-05	100.0%	8.7E-07	100.0%	4.1E-07
MT2A	17	23.5%	1	29.4%	1	41.2%	1	23.5%	1
CSF2	17	82.4%	0.0036	70.6%	0.1970	70.6%	0.1694	82.4%	0.0026
CCND1	15	93.3%	0.0002	53.3%	1	93.3%	0.0002	20.0%	1
APOE	15	66.7%	0.6166	66.7%	0.8173	26.7%	1	26.7%	1
TF	14	57.1%	1	42.9%	1	28.6%	1	35.7%	1
TP53	14	92.9%	0.0005	92.9%	0.0007	92.9%	0.0006	78.6%	0.0407
MYC	13	53.8%	1	61.5%	1	46.2%	1	53.8%	1
APOA2	13	23.1%	1	23.1%	1	53.8%	1	46.2%	1
APOB	12	41.7%	1	33.3%	1	8.3%	1	41.7%	1
CCNA2	12	33.3%	1	50.0%	1	50.0%	1	75.0%	0.2329
SFRS2	11	54.5%	1	63.6%	1	63.6%	1	54.5%	1
SFTPB	11	72.7%	0.6326	81.8%	0.1338	63.6%	1	54.5%	1
JUN	11	90.9%	0.0096	81.8%	0.1338	100.0%	0.0005	63.6%	1
ITGA2B	10	40.0%	1	60.0%	1	40.0%	1	70.0%	1
PLAU	10	100.0%	0.0013	100.0%	0.0018	90.0%	0.0310	100.0%	0.0010
HBZ	10	80.0%	0.2451	80.0%	0.3172	60.0%	1	80.0%	0.2072
IL6	10	90.0%	0.0260	90.0%	0.0354	80.0%	0.2839	80.0%	0.2072
CDC2	10	80.0%	0.2451	80.0%	0.3172	80.0%	0.2839	70.0%	1
NPPA	10	100.0%	0.0013	90.0%	0.0354	100.0%	0.0015	100.0%	0.0010
EPO	9	66.7%	1	66.7%	1	66.7%	1	55.6%	1
HMBS	9	44.4%	1	55.6%	1	55.6%	1	44.4%	1
MMP1	9	100.0%	0.0037	88.9%	0.0917	88.9%	0.0816	88.9%	0.0584
CDKN1A	8	62.5%	1	25.0%	1	75.0%	1	25.0%	1
INS	8	75.0%	1	62.5%	1	37.5%	1	62.5%	1
TRBC1	8	12.5%	1	25.0%	1	62.5%	1	0.0%	1
ADH1A	8	25.0%	1	12.5%	1	12.5%	1	12.5%	1
HBB	8	12.5%	1	12.5%	1	12.5%	1	12.5%	1
GH1	8	37.5%	1	50.0%	1	50.0%	1	37.5%	1
MBP	8	75.0%	1	62.5%	1	75.0%	1	62.5%	1
TYR	7	71.4%	1	71.4%	1	100.0%	0.0380	100.0%	0.0279

*Anzahl der TFBSs pro TF.

†Die Hintergrund-Konserviertheitsrate C_{seq}^{bg} beträgt für Mensch-Maus-Vergleiche 33.67% (KS $= 65\%$), für Mensch-Ratte-Vergleiche 34.91% (KS $= 64\%$), für Mensch-Kuh-Vergleiche 34.37% (KS $= 76\%$) und für Mensch-Hund-Vergleiche 32.89% (KS $= 78\%$).

‡Die Wahrscheinlicheit p, die beobachtete (oder eine größere) Differenz zwischen C_{seq} und C_{seq}^{bg} zufällig zu erhalten, wurde mit Gleichung (3.7) berechnet. Die p-Werte wurden Bonferroni-korrigiert.

4.6 Optimale Fenstergröße für den prädiktiven Ansatz

In den bisherigen Untersuchungen zur Sequenz-Konserviertheit (siehe 4.2, S. 69) war die Fenstergröße w zur Bestimmung von PI-Werten (siehe 3.3.5) jeweils durch die Länge der bekannten TFBSs im Alignment gegeben. Allerdings stellt sich für den Fall, dass im Alignment nach unbekannten TFBSs gesucht wird, d.h. beim sogenannten „prädiktiven Ansatz“, die Frage nach der optimalen Fenstergröße w zur Untersuchung eines paarweisen Alignments nach konservierten Regionen.

Um diese Frage zu beantworten, wurden 928 Mensch-TFBSs aus Datensatz IV in paarweisen LAGAN-Alignments (siehe 3.2.3.8, S. 37) zwischen Mensch und Maus mit verschiedenen Fenstergrößen zwischen 8 und 300 bp durchsucht (siehe 3.3.10, S. 46), um konservierte Bereiche zu bestimmen. Die Anteile der TFBSs und der Hintergrund-Sequenzen, die in diesen konservierten Bereichen liegen, wurden berechnet. Abbildung 4.22 zeigt die ROC-Kurven der erhaltenen Konserviertheitsraten für vier repräsentative Fenstergrößen.

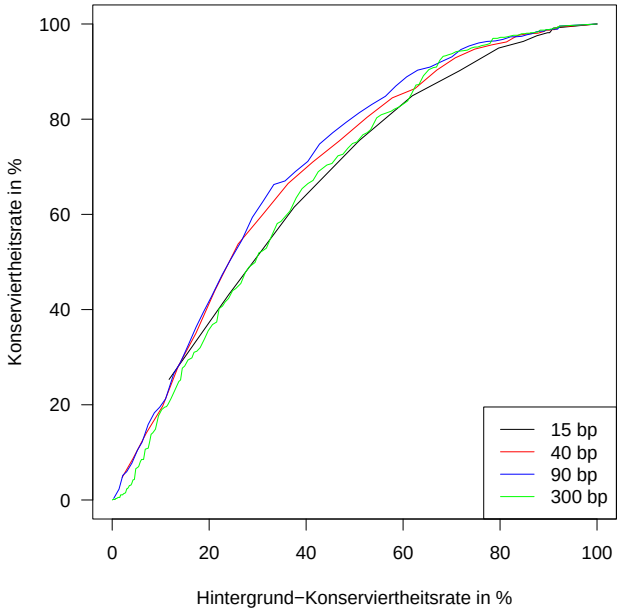

Abbildung 4.22: ROC-Diagramm für verschiedene Fenstergrößen. 928 Mensch-TFBSs und die Hintergrund-Sequenzen wurden mittels des prädiktiven Ansatzes auf ihre Konserviertheit hin in paarweisen LAGAN-Alignments zwischen Mensch und Maus untersucht. Die Abbildung zeigt die ROC-Kurven für vier repräsentative Fenstergrößen. Eine Fenstergröße von 90 bp ist kleineren und größeren Fenstern überlegen.

Für alle untersuchten Fenstergrößen wurde für die jeweilige ROC-Kurve der Punkt mit

dem minimalen Abstand zur linken oberen Ecke bestimmt. Eine Auftragung dieses minimalen Abstandes gegen die Fenstergröße w zeigt (siehe Abbildung 4.23), dass eine Fenstergröße von $w = 90$ bp am besten geeignet ist, um funktionelle TFBSs von Hintergrund-Sequenzen abzugrenzen. Der optimale KS für diese Fenstergröße beträgt 74%. Sowohl für kleinere als auch für größere Fenster steigt der minimale Abstand wieder an. Eine mögliche Erklärung dafür könnte sein, dass für kleine Fenstergrößen ($w < 90$ bp) die Wahrscheinlichkeit, ein Fenster fälschlicherweise als konserviert einzustufen, höher ist, wohingegen für große Fenstergrößen ($w > 90$ bp) die Wahrscheinlichkeit steigt, dass nicht-funktionelle Bereiche innerhalb des Fensters liegen und nicht mehr von funktionellen Bereichen getrennt werden können. Da die Genregulation in Eukaryoten auf dem Zusammenspiel und der Interaktion mehrerer TFBSs basiert, sollte die durchschnittliche Länge von funktionellen Regionen deutlich größer sein als die durchschnittliche Länge einer einzelnen TFBS von ca. 6-15 bp. Es ist denkbar, dass die optimale Fenstergröße von 90 bp der durchschnittlichen Länge einer funktionellen Region nahe kommt.

Auf der Webseite http://intergenomics.bioinf.med.uni-goettingen.de/seqfootprint.html existiert eine Oberfläche, mittels der Sequenzen aligniert und konservierte Fenster des Alignments visualisiert werden können.

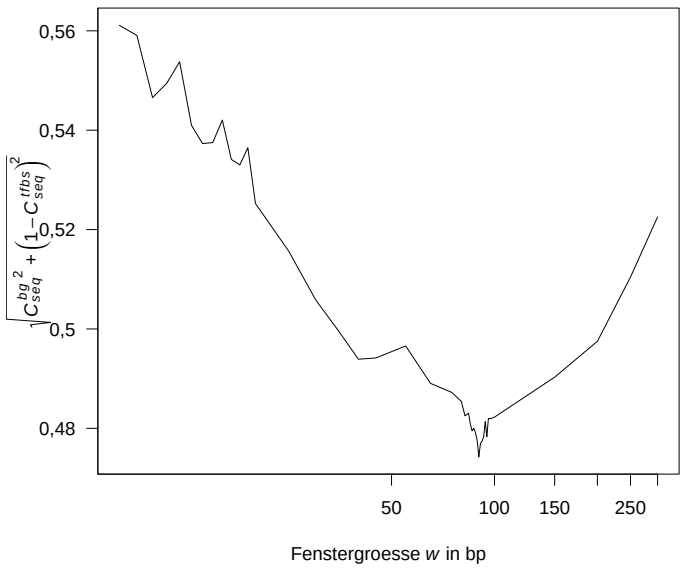

Abbildung 4.23: Für Fenstergrößen zwischen 8 und 300 bp wurde für jeden Punkt der ROC-Kurve der minimale Abstand zur linken oberen Ecke des ROC-Diagramms berechnet (siehe Abbildung 4.22). Je kleiner der Abstand, um so besser ist der Kompromiss zwischen der Konserviertheitsrate und falsch positiven Vorhersagen. Für eine Fenstergröße $w = 90$bp wird dieser Abstand minimal.

4.7 Anwendung des Hidden-Markov-Modells zur Vorhersage von Transkriptionsfaktor-Bindestellen

In den vorigen Kapiteln wurden bekannte TFBSs auf ihre Konserviertheit hin untersucht. Die erhaltenen Information über das Verhalten dieser TFBSs im Spezies-Vergleich wurden verwendet, um die Vorhersage von TFBSs zu verbessern. Die gängigste Methode zur Vorhersage von TFBSs ist der Vergleich von Sequenzen mit PSSMs (siehe 2.2.2, S. 19). Phylogenetisches Footprinting ist eine davon unabhängige Methode, die geeignet ist, funktionelle Bereiche, die TFBSs enthalten, in Promotor-Sequenzen zu identifizieren (siehe 4.2, S. 69).

Die Kombination beider Methoden sollte eine Verbesserung der Vorhersagequalität mit sich bringen. Eine Möglichkeit der Kombination der Vorhersagen von PSSMs und phylogenetischem Footprinting ist, die Suche mit PSSMs auf Sequenzabschnitte einzuschränken, die eine gewisse Sequenz-Konserviertheit überschreiten. Diese Sequenzabschnitte können mittels des prädiktiven Ansatzes (siehe 3.3.10, S. 46) bestimmt werden. Diese Art der Suche nach TFBSs erfordert die Wahl von zwei Schwellenwerten für den MSS und den PI-Wert. Weiterhin ist eine Linearkombination von MSS und PI-Wert möglich, LS genannt (siehe 3.4.7, S. 60), wodurch nur ein Schwellenwert gewählt werden muss. Beide Methoden haben den Nachteil, dass im Vorfeld entschieden werden muss, welchen Beitrag die Profil-Information der PSSM bzw. die Konserviertheit im Spezies-Vergleich zur Vorhersage von TFBSs leisten sollen. Um dies zu umgehen, wurde ein HMM entworfen (siehe 3.4, S. 48), das beide Informationsquellen nutzt, um auf einer DNA-Sequenz TFBSs vorherzusagen. Das HMM bedarf keiner vorherigen Entscheidung, welche Informationsquelle bevorzugt zu nutzen ist.

In Abschnitt 4.7.1 werden die Ergebnisse der empirischen Bestimmung der HMM-Parameter vorgestellt. In Abschnitt 4.7.2 werden die Ergebnisse der Vorhersage von TFBSs mittels des HMM, PSSMs und der Kombination von PSSMs und phylogenetischem Footprinting verglichen.

4.7.1 Bestimmung der HMM-Parameter

Die Parameter, die das Emissionsverhalten der Zustände **F** und **NB** des HMM charakterisieren, wurden empirisch für AVID-Alignments der OSPs aus Datensatz I bestimmt (siehe 3.4.4, S. 55). Tabelle 4.7 zeigt die Ergebnisse. Die Wahrscheinlichkeit, Lücken zu emittieren, ist für den Zustand **NB** mehr als doppelt so hoch wie für den Zustand **F**. Zudem ist die Wahrscheinlichkeit, dass ein emittiertes Nukleotid konserviert ist, für den Zustand **F** höher

als für den Zustand **NB**.

Tabelle 4.7: Empirisch ermittelte Parameterwerte der Zustände **F** und **NB** des HMM

Parameter	Wert
c_{f}	0.776
g_{f}	0.059
c_{nb}	0.599
g_{nb}	0.126

Tabelle 4.8 zeigt eine Übersicht der für die PSSMs aus Tabelle 3.7 empirisch ermittelten HMM-Parameter c_{tfbs} und g. Diese wurden je PSSM anhand aller Mensch-TFBSs aus Datensatz I bestimmt (siehe 3.4.4, S. 55). Die Wahrscheinlickeit c_{tfbs}, dass ein aus den Zuständen $\mathbf{B}_{\mathbf{1}}^{+}, \mathbf{B}_{\mathbf{2}}^{+}, \ldots, \mathbf{B}_{\lambda}^{+}$ bzw. $\mathbf{B}_{\mathbf{1}}^{-}, \mathbf{B}_{\mathbf{2}}^{-}, \ldots, \mathbf{B}_{\lambda}^{-}$ des HMM emittiertes Nukleotid konserviert ist, ist für alle untersuchten PSSMs größer als die Wahrscheinlichkeit c_{nb}, dass ein aus dem Zustand **NB** emittiertes Nukleotid konserviert ist. D.h. die TFBSs heben sich in ihrer Konserviertheit vom Hintergrund ab.

Tabelle 4.8: Empirisch ermittelte Parameterwerte der Zustände $\mathbf{B}_{\mathbf{1}}^{+}$, $\mathbf{B}_{\mathbf{2}}^{+}$, ..., $\mathbf{B}_{\lambda}^{+}$ bzw. $\mathbf{B}_{\mathbf{1}}^{-}$, $\mathbf{B}_{\mathbf{2}}^{-}$, ..., $\mathbf{B}_{\lambda}^{-}$ des HMM für verschiedene PSSMs

PSSM[‡]	TF	$N^{\dagger}$	$\lambda^{\ddagger}$	g	c_{tfbs}	$m_{w=15}$	$m_{w=90}$
M00761	p53	20	10	0.077	0.670	-2.740	-3.607
M00789	GATA	31	7	0.005	0.618	127.7	-639.8
M00912	CEBP	25	12	0.080	0.687	-1.928	-2.725
M00920	E2F	17	12	0.026	0.814	-2.098	-4.386
M00926	AP-1	33	8	0.038	0.886	-2.056	-2.952
M00931	Sp1	83	10	0.066	0.733	-4.160	-5.115
M00971	Ets	18	8	0.000	0.854	-1.778	-1.896
M00976	AHRHIF	17	9	0.087	0.745	-3.198	-3.891
M00981	CREBATF	18	9	0.034	0.870	-2.305	-4.611
M01031	HNF4	24	14	0.074	0.845	-1.234	-2.384
M01034	Ebox	24	10	0.014	0.913	-2.232	-3.136

Weiterhin wurde für jede PSSM der Wert m zur Berechnung des LS mit Gleichung (3.28) bestimmt. Dazu wurde allen Nukleotiden der OSPs, die TFBS der jeweiligen PSSM enthalten, mit dem prädiktiven Ansatz ein PI-Wert und mittels MATCH™ ein MSS zuge-

*TRANSFAC®-Zugriffsnummer der PSSM.
†Anzahl für den Menschen annotierter TFBSs.
‡Länge der PSSM.

ordnet. Zur Bestimmung der PI-Werte wurden zwei Fenstergrößen eingesetzt: Zum einen die in Abschnitt 4.6 (S. 102) ermittelte optimale Fenstergröße von $w = 90$ bp, zum anderen die durchschnittliche Länge der TFBSs aus Datensatz I von $w = 15$ bp. Je größer dabei der erhaltene absolute Wert für m ist, um so größer ist der Beitrag des MSS zum LS. Für die 31 TFBSs der PSSM mit der TRANSFAC®-Zugriffsnummer M00789 ergaben sich sehr hohe absolute Werte für m. D.h. in diesem Fall wird der Beitrag des PI-Wertes zum LS in Gleichung (3.27) vernachlässigbar.

4.7.2 Vergleich der Vorhersagen von MATCH™ und des HMM

MATCH™, eine Kombination aus MATCH™ und phylogenetischem Footprinting sowie das HMM wurden in ihrer Fähigkeit, TFBSs zu identifizieren, verglichen, indem überprüft wurde, wieviele Nukleotide der bekannten TFBSs und wieviele Nukleotide des Hintergrundes für eine bestimmte PSSM vorhergesagt wurden (siehe 3.4.7, S. 60). Dazu wurden für jede PSSM aus Tabelle 4.8 alle OSPs, die zur Konstruktion der PSSM benutzte Mensch-TFBS enthalten, mit dem HMM bzw. MATCH™ durchsucht. Das HMM wurde mit den Parameterwerten aus Tabelle 4.8 initialisiert, wobei der Parameter f variiert wurde. Der Parameter $f \in [0, 1]$ beeinflußt die Übergangswahrscheinlichkeit, die Zustände **F** bzw. **NB** des HMM zu betreten. Für die Grenzfälle $f = 0$ bzw. $f = 1$ sind die Zustände **F** bzw. **NB** des HMM nicht mehr erreichbar. Als Eingabe für das HMM dienten die alignierten OSPs, die PSSM und verschiedene Werte für den Parameter p (siehe 3.4.5, S. 56), um Sensitivitätswerte zwischen 0 und 100% zu erhalten. Auch die Schwellenwerte für MSS, PI-Wert und LS wurden variiert, um für die Vorhersagen von MATCH™ oder einer Kombination aus MATCH™ und phylogenetischem Footprinting Sensitivitätswerte zwischen 0 und 100% abzudecken. Für jede Methode und jeden Wert der Sensitivität wurde der jeweilige positive Vohersagewert berechnet (siehe 3.4.7, S. 60). Eine Methode zur Vorhersage von TFBSs ist einer anderen überlegen, wenn sie bei gleicher Sensitivität einen höheren positiven Vorhersagewert aufweist.

Die Abbildungen 4.24 bis 4.34 zeigen die für die jeweilige PSSM mit den einzelnen Methoden erhaltenen Sensitivitäten und positiven Vorhersagewerte. Für die Kombination von MATCH™ und phylogenetischem Footprinting, d.h. für die Eingrenzung der Vorhersagen mit MATCH™ auf konservierte Bereiche oder für die Linearkombination aus MSS und PI-Wert, wurde nur die jeweils beste Kurve in die Abbildung eingetragen. Für das HMM wurden zwei Kurven eingetragen, die auf verschiedenen Werten des Parameters f basieren: Zum einen wurde $f = 0$ gesetzt und zum anderen ein Wert von $f > 0$ gewählt, der eine verbesserte Vorhersage lieferte.

Für acht der elf untersuchten PSSMs führte eine Eingrenzung der Vorhersagen von MATCH™ auf konservierte Bereiche (M00789, M00912, M01031 und M01034; siehe Abbildungen 4.24, 4.25, 4.30 und 4.31) bzw. eine Linearkombination des MSS und des PI-Wertes (M00926, M00931, M00971 und M00976; siehe Abbildungen 4.26, 4.27, 4.28 und 4.29) bei gleicher Sensitivität zu erhöhten positiven Vohersagewerten gegenüber den Vorhersagen von MATCH™. Es war keine Tendenz zu erkennen, ob die Wahl eines Schwellenwertes für den LS oder die Wahl von zwei Schwellenwerten für den PI-Wert und den MSS generell zu bevorzugen ist. Eine Fenstergröße von $w = 90$ bp lieferte in den meisten Fällen bessere Ergebnisse (siehe dazu auch 4.6, S. 102). Nur für die PSSM M00926 wurde mit einer Linearkombination aus PI-Wert für eine Fenstergröße von $w = 15$ bp und MSS eine erhebliche Verbesserung erzielt.

Die Vorhersagen des HMM ergaben für die o.g. acht PSSMs (M00789, M00912, M00926, M00931, M00971, M00976, M01031 und M01034) bei gleicher Sensitivität höhere positive Vorhersagewerte als die Vorhersagen von MATCH™ oder einer Kombination von MATCH™ und phylogenetischem Footprinting (siehe Abbildungen 4.24, 4.25, 4.26, 4.27, 4.28, 4.29, 4.30 und 4.31). Besonders deutlich zeigte sich dies im Bereich von Sensitivitätswerten zwischen 40 und 75%. Beispielsweise machte MATCH™ für die PSSM M00926 (AP-1) bei 25 erkannten TFBSs (Sensitivität von 75%) 82 zusätzliche Vorhersagen, während mit dem HMM bei gleicher Sensitivität nur 19 zusätzliche TFBSs vorhergesagt wurden. Dies entspricht einer Reduzierung der falsch positiven Vorhersagen auf weniger als ein Viertel. Tabelle 4.9 zeigt eine Übersicht der für eine Sensitivität von 60% erhaltenen positiven Vorhersagewerte.

Tabelle 4.9: Erhaltene positive Vohersagewerte für das HMM und MATCH™ bei einer Sensitivität von 60%

PSSM	TF	PPV_{HMM}	f	$PPV_{MATCH^{TM}}$
M00761	p53	30.7%	0	24.8%
M00789	GATA	54.5%	0.5	44.2%
M00912	CEBP	55.6%	0	27.4%
M00920	E2F	83.3%	1	81.9%
M00926	AP-1	57.6%	1	34.7%
M00931	Sp1	31.8%	0.5	23.1%
M00971	Ets	15.6%	1	8.0%
M00976	AHR/HIF	43.5%	0	25.6%
M00981	CREB/ATF	51.3%	1	67.3%
M01031	HNF4	73.7%	1	27.4%
M01034	Ebox	60.8%	1	47.0%

Nur für die PSSM mit der TRANSFAC®-Zugriffsnummer M00981 wurde bei einer Sensitivität von 60% ein niedrigerer positiver Vohersagewert für die Vorhersagen des HMM als für die von MATCH™ erhalten. Für fünf der untersuchten PSSMs (M00912, M00926, M00931, M00976, M01031) war das HMM in seinen Vorhersagen einer Kombination aus den Vorhersagen von MATCH™ und phylogenetischem Footprinting deutlich überlegen, d.h. in diesen Fällen wurden die beiden Informationsquellen zur Vorhersage einer TFBSs, also das Profil der PSSM und die Konserviertheit funktioneller Bereiche, optimal verknüpft. Beispielsweise führte für die PSSM M01031 bei einer Sensitivität von 60% eine Eingrenzung der Vorhersagen von MATCH™ auf konservierte Bereiche, die mit einer Fenstergröße von $w = 90$ bp bestimmt wurden, zu einem positiven Vorhersagewert von 39%, während die Vorhersagen des HMM einen positiven Vorhersagewert von 74% erreichten (siehe Abbildung 4.30).

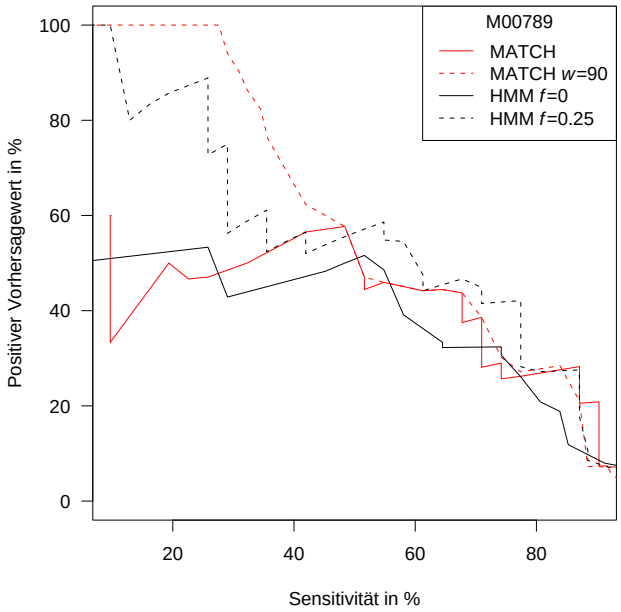

Abbildung 4.24: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00789 (GATA). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Beschränkung der Vorhersagen von MATCH™ auf konservierte Fenster, die mit einer Fenstergröße von $w = 90$ bp bestimmt wurden, erhält man für geringe Sensitivitätswerte höhere positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$. Durch ein Anheben des Parameters f auf 0.25 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve), die den Vorhersagen von MATCH™ überlegen ist.

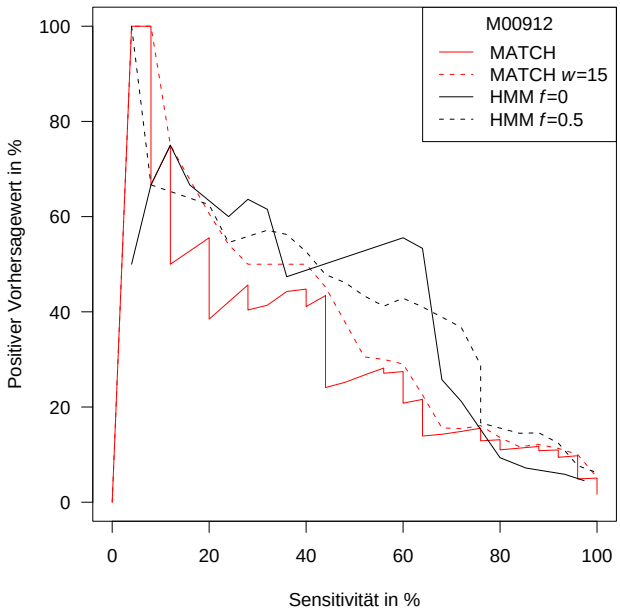

Abbildung 4.25: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00912 (C/EBP). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Beschränkung der Vorhersagen von MATCH™ auf konservierte Fenster, die mit einer Fenstergröße von $w = 15$ bp bestimmt wurden, erhält man generell höhere positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die bis zu einer Sensitivität von 70% den Vorhersagen von MATCH™ überlegen sind. Durch ein Anheben des Parameters f auf 0.5 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte $\geq 65\%$.

Für die PSSMs M00761, M00920 und M00981 brachte das HMM gegenüber MATCH™ keine oder nur geringe Vorteile in der Vorhersagequalität (siehe Abbildungen 4.32, 4.33, 4.34). Dies kann auf unterschiedliche Gründe zurückgeführt werden. Für die PSSM M00920 ergaben die Vorhersagen mit MATCH™ für Sensitivitätswerte bis ca. 70% bereits recht hohe positive Vorhersagewerte (PPV $> 75\%$; siehe Abbildung 4.33). In diesem Fall wurden durch eine Kombination der Vorhersagen von MATCH™ und phylogenetischem Footprinting oder für die Vorhersagen des HMM keine höheren positiven Vorhersagewerte erhalten. Eine ähnliche Beobachtung wurde für die PSSM M00981 gemacht, für die die Vorhersagen von MATCH™ bis zu einer Sensitivität von 65% positive Vorhersagewerte von ca. 60% aufwiesen (siehe Abbildung 4.34). Hier ergaben die Vorhersagen des HMM für $f = 1$ nur ab einer Sensitivität von 70% höhere positive Vorhersagewerte als die Vorhersagen von MATCH™. Die TFBSs der PSSMs M00920 bzw. M00981 sind zwar gut

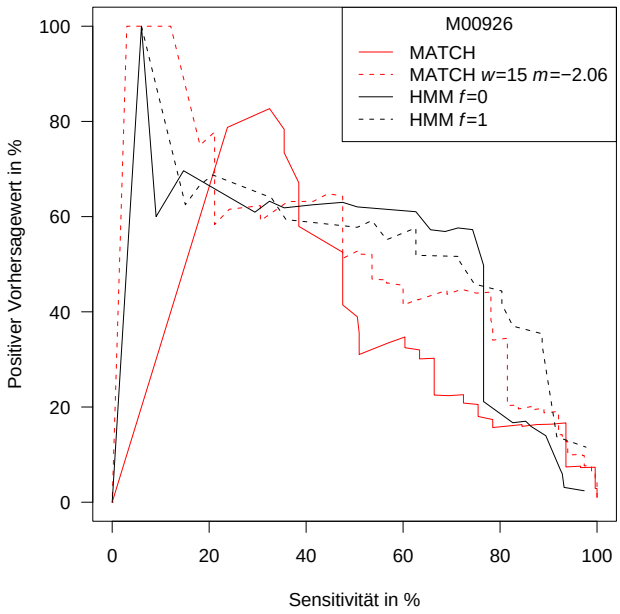

Abbildung 4.26: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00926 (AP-1). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Linearkombination ($m = -2.06$) der MSS-Werte mit den mittels einer Fenstergröße von $w = 15$ bp erhaltenen PI-Werten, erhält man ab einer Sensitivität von 40% höhere positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die bis zu einer Sensitivität von 75% den Vorhersagen von MATCH™ überlegen sind. Durch ein Anheben des Parameters f auf 1 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte $\geq 75\%$.

konserviert ($c_{\text{tfbs}} = 0.814$ bzw. 0.870), aber in diesen Fällen scheint das Muster der PSSM und nicht die Konserviertheit der TFBSs ausschlaggebend für die Vorhersage von TFBSs zu sein, sodass das HMM hier keine Vorteile gegenüber MATCH™ aufzeigte. Die TFBSs der PSSM M00761 sind im Gegensatz dazu nur schwach konserviert ($c_{\text{tfbs}} = 0.67$), wodurch der Einfluss der Konserviertheit auf die Vorhersagen des HMM vermutlich unbedeutend war.

Einfluss des Parameters f

Für einige der untersuchten PSSMs (M00920, M00926, M01031 und M01034; siehe Abbildungen 4.33, 4.26, 4.30 und 4.31) war für die Vorhersagen des HMM zu beobachten, dass der positive Vorhersagewert für Sensitivitätswerte oberhalb von 60% rapide abnahm,

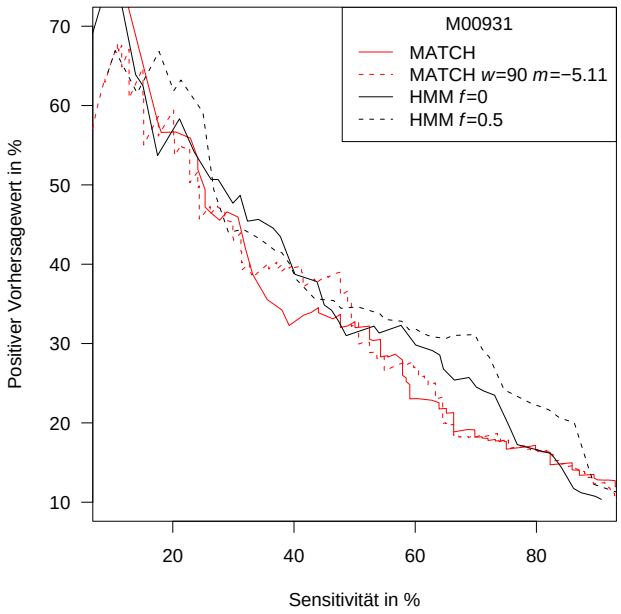

Abbildung 4.27: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00931 (Sp1). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Linearkombination ($m = -5.11$) der MSS-Werte mit den mittels einer Fenstergröße von $w = 90$ bp erhaltenen PI-Werten, erhält man für Sensitivitätswerte zwischen 30 und 50% leicht verbesserte positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die bis zu einer Sensitivität von 80% den Vorhersagen von MATCH™ überlegen sind. Durch ein Anheben des Parameters f auf 0.5 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte $\geq 60\%$.

wenn der Parameter $f = 0$ gesetzt war. In diesem Fall war der Zustand **F** des HMM nicht aktiv. Dieser Zustand wurde eingeführt, um Sequenzabschnitte, die hoch konserviert sind, aber nicht dem Profil der PSSM ähneln, zu emittieren (siehe 3.4.2, S. 51). Wurde für die o.g. PSSMs der Parameter $f = 1$ gesetzt, wurden für hohe Sensitivitätswerte höhere positive Vorhersagewerte erhalten. Das HMM machte in diesem Fall weniger Vorhersagen, die zwar hoch konserviert sind, aber nur eine mäßige Ähnlichkeit zum Profil der PSSM aufweisen: Für die PSSM M00926 wurden beispielhaft die Vorhersagen des HMM für die gewählten Parameter $f = 0$ und $f = 1$ bei einer Sensitivität von 82.6% verglichen. Für $f = 0$ wurden 161 Vorhersagen gemacht, für $f = 1$ hingegen nur 71. Von diesen waren 64 Vorhersagen identisch. Die 97 Vorhersagen, die nur für $f = 0$ gemacht wurden, waren stark konserviert (durchschnittlicher PI-Wert von 96%), aber sie zeigten nur eine moderate

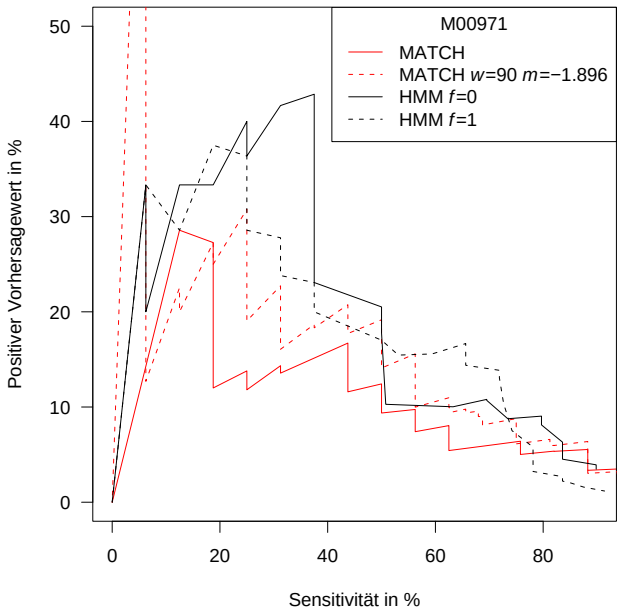

Abbildung 4.28: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00971 (Ets). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Linearkombination ($m = -1.896$) der MSS-Werte mit den mittels einer Fenstergröße von $w = 90$ bp erhaltenen PI-Werten, erhält man für Sensitivitätswerte $\geq 20\%$ verbesserte positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die bis zu einer Sensitivität von 50% den Vorhersagen von MATCH™ überlegen sind. Durch ein Anheben des Parameters f auf 1 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte zwischen 50 und 70%.

Ähnlichkeit zur PSSM (durchschnittlicher MSS von 0.794). Für die 33 annotierten TFBSs wurde ein durchschnittlicher PI-Wert von 89% und ein deutlich höherer durchschnittlicher MSS von 0.937 erhalten. Diesen Werten kamen die Vorhersagen für $f = 1$ deutlich näher. Sie wiesen einen durchschnittlichen PI-Wert von 92% und einen durchschnittlichen MSS von 0.919 auf. Diese Daten zeigen, dass für den Fall $f = 0$ bei hohen Sensitivitätswerten die Konserviertheit das Profil der PSSM als Informationsquelle, eine TFBS vorherzusagen, überwiegt, d.h. für $f = 0$ werden konservierte, aber dem Profil der PSSM nur mäßig ähnelnde Sequenzabschnitte mit höherer Wahrscheinlichkeit in den Zuständen $\mathbf{B}_1^+, \mathbf{B}_2^+, \ldots, \mathbf{B}_\lambda^+$ bzw. $\mathbf{B}_1^-, \mathbf{B}_2^-, \ldots, \mathbf{B}_\lambda^-$ als im Zustand **NB** emittiert. Für PSSMs, die eine hohe Wahrscheinlichkeit aufweisen, dass ein in den Zuständen $\mathbf{B}_1^+, \mathbf{B}_2^+, \ldots, \mathbf{B}_\lambda^+$ bzw. $\mathbf{B}_1^-, \mathbf{B}_2^-, \ldots, \mathbf{B}_\lambda^-$ emittiertes Nukleotid konserviert ist, z.B. $c_{\text{tfbs}} \geq 0.8$, ist der Zustand **F** besser als der Zustand **NB**

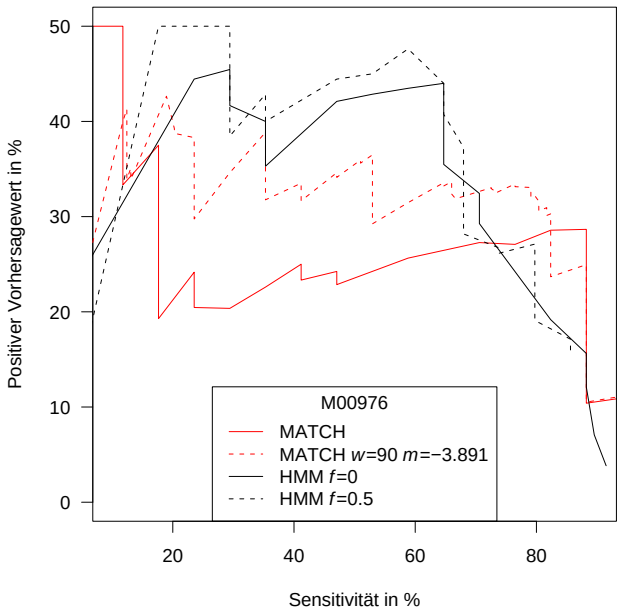

Abbildung 4.29: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00976 (AHR/HIF). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Linearkombination ($m = -3.891$) der MSS-Werte mit den mittels einer Fenstergröße von $w = 90$ bp erhaltenen PI-Werten, erhält man für Sensitivitätswerte zwischen 20 und 80% verbesserte positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die bis zu einer Sensitivität von 70% den Vorhersagen von MATCH™ überlegen sind. Durch ein Anheben des Parameters f auf 0.5 erhält man eine geringfügige Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte zwischen 20 und 60%.

geeignet, um die gesuchten TFBSs von ihrer Umgebung zu diskriminieren. Der Zustand **F** besitzt eine höhere Wahrscheinlichkeit, dass ein in diesem Zustand emittiertes Nukleotid konserviert ist ($c_{\mathrm{f}} = 0.776$), als der Zustand **NB** ($c_{\mathrm{nb}} = 0.599$). Wenn der Zustand **F** aktiv ist, werden daher konservierte, aber dem Profil der PSSM nur mäßig ähnelnde Sequenzabschnitte mit höherer Wahrscheinlichkeit im Zustand **F** als in den Zuständen $\mathbf{B}_1^+, \mathbf{B}_2^+, \ldots, \mathbf{B}_\lambda^+$ bzw. $\mathbf{B}_1^-, \mathbf{B}_2^-, \ldots, \mathbf{B}_\lambda^-$ emittiert.

Wenn die Wahrscheinlichkeit c_{tfbs}, dass ein aus den Zuständen $\mathbf{B}_1^+$, $\mathbf{B}_2^+$,..., $\mathbf{B}_\lambda^+$ bzw. $\mathbf{B}_1^-$, $\mathbf{B}_2^-$,..., $\mathbf{B}_\lambda^-$ emittiertes Nukleotid konserviert ist, zwischen 0.62 und 0.75 liegt (M00789, M00912, M00931, M00976), war es von Vorteil, wenn sowohl der Zustand **NB** als auch der Zustand **F** des HMM aktiv waren. Hier wurde mit Werten für f zwischen 0.25 und 0.5 eine bessere Diskrimination zwischen den annotierten TFBSs und schwach oder hoch

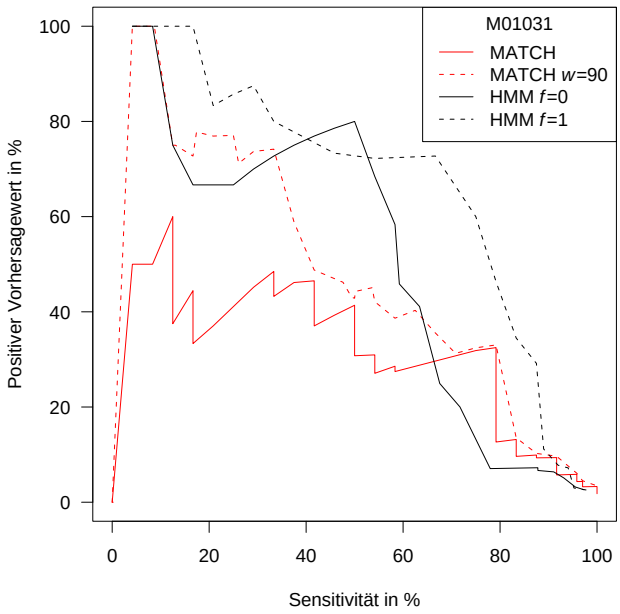

Abbildung 4.30: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M01031 (HNF4). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Beschränkung der Vorhersagen von MATCH™ auf konservierte Fenster, die mit einer Fenstergröße von $w = 15$ bp bestimmt wurden, erhält man generell höhere positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die bis zu einer Sensitivität von 60% den Vorhersagen von MATCH™ überlegen sind. Durch ein Anheben des Parameters f auf 1 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) über den gesamten Wertebereich der Sensitivität.

konservierten Bereichen, die keine TFBSs des gesuchten Typ enthalten, erreicht.

Fazit und Ausblick

Zusammenfassend lässt sich sagen, dass die Vorhersagen des HMM denen von MATCH™ überlegen oder zumindest ebenbürtig sind. In den Fällen, in denen eine Kombination der Vorhersagen von MATCH™ und phylogenetischen Footprinting Vorteile bei der Vorhersage brachte, wurden diese bei den Vorhersagen des HMM noch ausgeprägter. Zudem zeigte sich, dass mit steigender Konserviertheit der untersuchten TFBSs auch die Wahrscheinlichkeit, den Zustand **F** des HMM zu betreten, steigen sollte, damit konservierte, aber dem Profil der PSSM nur mäßig ähnelnde Sequenzabschnitte aus diesem Zustand emittiert und nicht als TFBSs vom HMM ausgegeben werden.

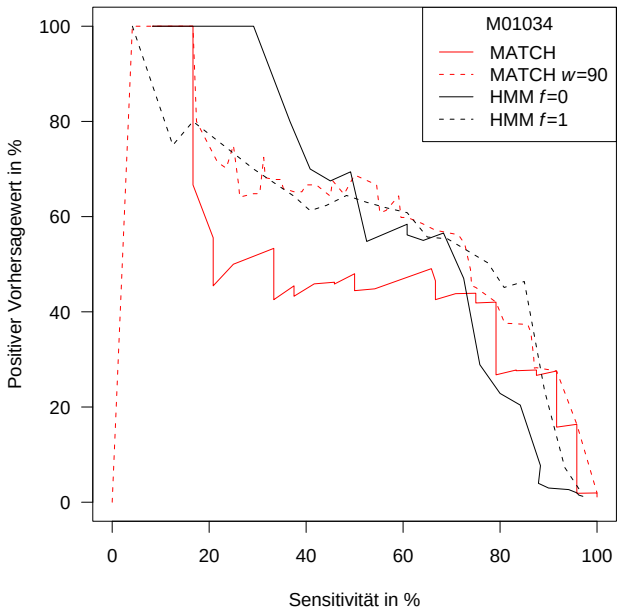

Abbildung 4.31: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M01034 (Ebox). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Beschränkung der Vorhersagen von MATCH™ auf konservierte Fenster, die mit einer Fenstergröße von $w = 90$ bp bestimmt wurden, erhält man generell höhere positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die bis zu einer Sensitivität von 45% den Vorhersagen von MATCH™ überlegen sind. Durch ein Anheben des Parameters f auf 1 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte $\geq 45\%$.

Für den Fall, dass die Zustände **F** oder **NB** beide erreichbar sind (d.h. $0 < f < 1$), ist anzumerken, dass der Viterbi-Pfad möglicherweise nicht optimal für die Vorhersage von TFBSs ist. Es ist in diesem Fall nicht von Bedeutung, ob sich das HMM an bestimmten Positionen der untersuchten Sequenz im Zustand **F** oder **NB** befunden hat, sondern nur, ob sich das HMM in den Zuständen $\mathbf{B}_1^+, \mathbf{B}_2^+, \ldots, \mathbf{B}_\lambda^+$ bzw. $\mathbf{B}_1^-, \mathbf{B}_2^-, \ldots, \mathbf{B}_\lambda^-$ befunden hat oder nicht. Daher ist es möglich, dass es viele verschiedene Pfade durch das Modell gibt, die dieselbe Vorhersage ergeben und in ihrer Gesamtheit eine höhere Wahrscheinlichkeit als der Viterbi-Pfad aufweisen würden. Eine Möglichkeit, um ähnliche Pfade zu erkennen, ist, Zuständen, die die gleiche Sequenzeigenschaft repräsentieren, eine gemeinsame Kennzeichnung (engl. „label“) zu geben. Anstatt den wahrscheinlichsten Zustandspfad zu berechnen, kann das wahrscheinlichste „labelling“ der untersuchten Sequenz berechnet werden. Eine signifikante Verbesserung der Vorhersagequalität ist durch dieses Vorgehen aber nicht zu

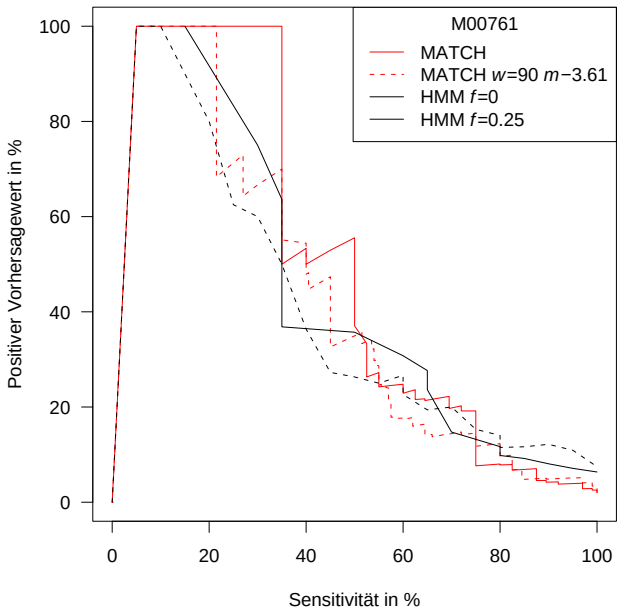

Abbildung 4.32: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00761 (p53). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Linearkombination ($m = -3.61$) der MSS-Werte mit den mittels einer Fenstergröße von $w = 90$ bp erhaltenen PI-Werten, erhält man generell niedrigere positive Vorhersagewerte (rot gestrichelte Kurve). Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die den Vorhersagen von MATCH™ nicht überlegen sind. Durch ein Anheben des Parameters f auf 0.25 erhält man eine geringfügige Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte $\geq 70\%$.

erwarten.

Im Anwendungsfall stellt sich für den Nutzer die Frage, mit welchen Werten für p und f das Programm initialisiert werden soll. Für die untersuchten PSSMs ist es möglich, für jede dieser PSSMs tabellierte Werte für p und f zur Verfügung zu stellen, die bestimmte Sensitivitäten und positive Vorhersagewerte abdecken. Um die Benutzerfreundlichkeit allerdings zu erhöhen und eine Skalierbarkeit zu gewährleisten, ist es denkbar, dem Programm einen Sensitivitätsparameter zu übergeben und über eine Teststatistik, die auf den Inhaltsmodellen der Zustände **NB**, **F** und $\mathbf{B}_1^+, \mathbf{B}_2^+, \ldots, \mathbf{B}_\lambda^+$ bzw. $\mathbf{B}_1^-, \mathbf{B}_2^-, \ldots, \mathbf{B}_\lambda^-$ basiert, die Übergangswahrscheinlichkeiten des HMMs so berechnen zu lassen, dass die geforderte Sensitivität erzielt wird. In diesem Fall könnte das HMM auch auf PSSMs, die nicht untersucht wurden, angewandt werden, wobei als einziges Vorwissen Schätzwerte für die Parameter g und c_{tfbs} nötig wären. Wenn es nicht möglich ist, diese Parameter anhand der

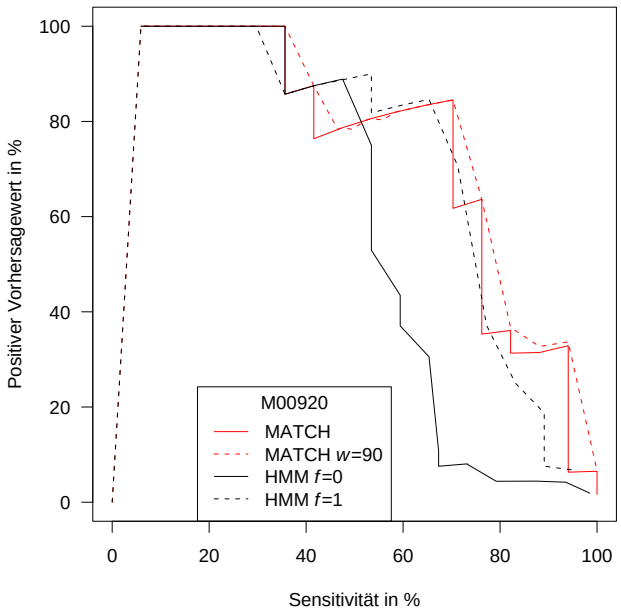

Abbildung 4.33: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00920 (E2F). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Beschränkung der Vorhersagen von MATCH™ auf konservierte Fenster, die mit einer Fenstergröße von $w = 90$ bp bestimmt wurden (rot gestrichelte Kurve), erhält man geringfügig höhere positive Vorhersagewerte. Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die den Vorhersagen von MATCH™ unterlegen sind. Durch ein Anheben des Parameters f auf 1 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve), die sich den mittels MATCH™ erhaltenen Werten nahe kommt.

zur PSSM-Konstruktion benutzten TFBSs empirisch zu ermitteln, könnten alternativ die Durchschnittswerte für die Gesamtheit aller bekannten TFBSs, d.h. g_{f} (relativer Anteil aller Lücken in allen untersuchten TFBSs) und c_{f} (relativer Anteil konservierter Nukleotide in allen untersuchten TFBSs), verwendet werden.

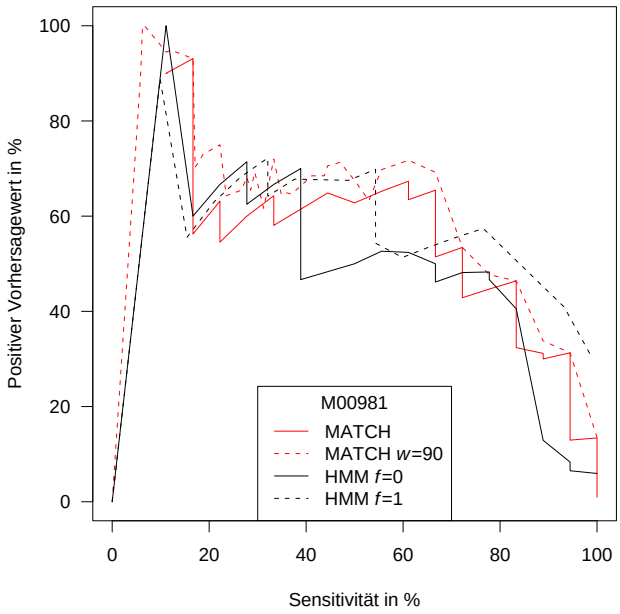

Abbildung 4.34: Vergleich der Vorhersagen des HMM und von MATCH™ für PSSM M00981 (CREB/ATF). Die Kurven zeigen die erhalten positiven Vorhersagewerte der einzelnen Methoden bei gegebener Sensitivität. Die rote Kurve zeigt die für MATCH™ erhaltenen Werte. Durch eine Beschränkung der Vorhersagen von MATCH™ auf konservierte Fenster, die mit einer Fenstergröße von $w = 90$ bp bestimmt wurden (rot gestrichelte Kurve), erhält man geringfügig höhere positive Vorhersagewerte. Die schwarze Kurve zeigt die für das HMM erhaltenen Werte für den Fall $f = 0$, die den Vorhersagen von MATCH™ unterlegen sind. Durch ein Anheben des Parameters f auf 1 erhält man eine Verbesserung der Vorhersagequalität (schwarz gestrichelte Kurve) für Sensitivitätswerte $\geq 70\%$.

Kapitel 5

Zusammenfassung

Im Rahmen dieser Arbeit wurde die Nützlichkeit des phylogenetischen Footprintings zur Vorhersage von TFBSs in nicht-codierenden Sequenzen untersucht, um darauf aufbauend die Vorhersage von TFBSs mithilfe eines HMM zu verbessern. Die zugrunde liegende Annahme des phylogenetischen Footprintings ist, dass regulatorische Bereiche unter einem selektiven Druck stehen und sich daher im Spezies-Vergleich durch hohe Sequenzähnlichkeiten von einem Hintergrund neutral divergierender Sequenz abheben.

Um dies zu überprüfen, wurde die Konserviertheit von 2578 experimentell bekannten TFBSs aus der TRANSFAC®-Datenbank zwischen Mensch und Maus bzw. Ratte untersucht. Für den Erfolg des phylogenetischen Footprintings ist die Sicherstellung der orthologen Beziehung zwischen den zu vergleichenden Sequenzen entscheidend. Es wurde gezeigt, dass orthologe Sequenzen nicht grundsätzlich anhand eines identischen relativen Abstands zur TSS in beiden Genomen lokalisiert werden können, da die annotierten TSSs orthologer Gene nicht zwangsläufig orthologe Positionen sind. Dies ist zumeist auf eine fehler- oder lückenhafte Annotation der TSS in einer oder beiden Spezies zurückzuführen. Im Rahmen dieser Arbeit wurde ein Verfahren entwickelt, dass diese Problematik umgeht, indem orthologe Sequenzen durch die Suche nach Sequenzhomologien in der Umgebung annotierter orthologer Gene lokalisiert werden. 25% von 775 OSPs aus Mensch und Maus oder Ratte, die mit diesem Verfahren lokalisiert wurden, unterschieden sich in ihrem Abstand zur annotierten TSS um mehr als 500 bp, 18% sogar um mehr als 1000 bp. Würde man Sequenzen einer Länge von 1000 bp stromaufwärts der annotierten TSS orthologer Gene miteinander vergleichen, wären diese Sequenzen somit in 18% der Fälle nicht ortholog zueinander.

Um zu überprüfen, ob sich annotierte TFBSs in paarweisen Alignments durch eine erhöhte Konserviertheit von der sie umgebenden Sequenz abheben, wurde die Wahl eines Schwellenwertes zur Bestimmung konservierter Bereiche unter zwei Gesichtspunkten ana-

lysiert: dem Anteil konservierter TFBSs an allen untersuchten TFBSs und dem Anteil konservierter Sequenzen an allen Sequenzen, die keine bzw. keine bekannte Funktion tragen. Ein KS von 65% lieferte dabei für Vergleiche zwischen Mensch und Nagetier eine optimale Trennung zwischen bekannten TFBSs und nicht-funktionellen Sequenzabschnitten. 72% der 2578 untersuchten TFBS und 35% der Hintergrund-Sequenzen sind bei diesem KS konserviert, d.h. die Funktionalität der TFBSs spiegelt sich deutlich in ihrer Konserviertheit wider.

Die Wahl des Alignment-Algorithmus hatte nur einen marginalen Einfluss auf die erhaltenen Ergebnisse, da Mensch- und Nagetier-Sequenzen eine genügend hohe Ähnlichkeit aufweisen, sodass die meisten Alignment-Programme ähnliche Ergebnisse produzieren.

Die Sequenz-Konserviertheit von TFBSs zeigt spezifische Unterschiede und variiert unter anderem in Abhängigkeit vom zugehörigen TF. Zum Beispiel sind TFBSs der TFs MyoD, MEF-2, SRF und NF-AT1 stark konserviert, wohingegen TFBSs der TFs Sp1, C/EBPα und AP-2αA nur schwach konserviert sind. Weiterhin unterscheiden sich die einzelnen Nukleotide von TFBSs in ihrer Konserviertheit. Die Nukleotide, die den größten Beitrag zur Spezifität einer bestimmten DNA-TF-Bindung leisten, sind meistens stärker konserviert als die übrigen Nukleotide einer TFBS. Deutliche Unterschiede in der Sequenz-Konserviertheit von TFBSs zeigen sich auch in Abhängigkeit von der Funktion des regulierten Gens: TFBSs von Genen, die in die Regulation der Transkription, die Entwicklung oder die Signaltransduktion involviert sind, zeigen höhere Konserviertheitsraten als TFBSs von Genen, die an katalytischen Vorgängen oder Transportprozessen beteiligt sind. Die Sequenz-Konserviertheit von TFBSs wird auch durch das Zusammenspiel und die Interaktion einzelner TFs beeinflusst. Die TRANSCompel®-Datenbank enthält Einträge von Paaren von TFBSs, deren TFs eine Protein-Protein-Wechselwirkung ausbilden. Die 405 untersuchten TFBSs der TRANSCompel®-Datenbank sind stärker konserviert als die 2578 untersuchten TFBSs der TRANSFAC®-Datenbank, was darauf hinweist, dass komplexe, zusammengesetzte regulatorische Elemente einem erhöhten evolutionären Druck unterliegen.

Die Sequenz-Konserviertheit einer TFBS ist nicht immer als alleiniges Kriterium ausreichend, um zu bestimmen, ob die TFBS auch in der orthologen Sequenz existiert. Da die DNA-Bindungs-Spezifität für TFs im Allgemeinen recht gering ist, kann die Sequenz einer TFBSs nur schwach konserviert sein und der TF dennoch an die orthologe Sequenz binden. Andererseits ist es möglich, dass eine TFBSs durch eine Mutation, Insertion oder Deletion an einer essentiellen Position zerstört wurde, die restliche Sequenz aber dennoch hoch konserviert ist. Aus diesem Grund wurde zusätzlich das Konzept der Muster-Konserviertheit eingeführt. Für einige TFs ergaben sich signifikante Unterschiede zwi-

schen den Konserviertheitsraten auf Sequenz- und Musterebene, z.B. sind 90% der AP-2αA TFBSs auf Musterebene konserviert, aber nur 48% auf Sequenzebene. Das Konzept der Muster-Konserviertheit ist daher für TFs nützlich, bei denen rein Sequenz-basiertes phylogenetisches Footprinting versagen könnte.

Da die Anzahl sequenzierter eukaryotischer Genome in den letzten Jahren gestiegen ist, stehen auch andere Spezies als Maus oder Ratte zur Verfügung, um funktionelle Bereiche im Humangenom zu detektieren. Im Rahmen dieser Arbeit wurden für einen Datensatz von 928 Mensch-TFBSs aus der TRANSFAC®-Datenbank paarweise Vergleiche zwischen Mensch und Maus, Ratte, Hund oder Kuh durchgeführt. Paarweise Vergleiche zwischen Mensch und Maus oder Ratte erwiesen sich denen zwischen Mensch und Hund oder Kuh als überlegen. Für Hund und Kuh ist der evolutionäre Abstand zum Menschen geringer, wodurch konservierte Bereiche in einem Alignment häufiger auf die gemeinsame Abstammung als auf evolutionären Druck zurückzuführen sind. Die Konserviertheit der untersuchten TFBSs ist für die einzelnen Speziesvergleichen insgesamt positiv korreliert, für einen Teil der untersuchten TFBSs wurden allerdings Spezies-spezifische Unterschiede in der Konserviertheit festgestellt, welche vermutlich auf eine unterschiedliche Regulation bestimmter Gene in den einzelnen Spezies zurückzuführen sind.

Ein Ziel dieser Arbeit war, die Vorhersage von TFBSs mittels der aus Speziesvergleichen erhaltenen Informationen zu verbessern. Im Anwendungsfall stellt sich bei der Suche nach unbekannten TFBSs die Frage, wie man optimaler Weise konservierte Bereiche in einem Alignment bestimmt. Dazu wurden konservierte Bereiche in den untersuchten Alignments mit unterschiedlichen Fenstergrößen bestimmt: Eine Fenstergröße von 90 bp war kleineren und größeren Fenstergrößen überlegen und lieferte in Kombination mit einem KS von 74% eine optimale Diskrimination zwischen annotierten TFBSs und Hintergrund-Sequenzen. Eine gängige Methode zur Vorhersage von TFBSs ist der Vergleich von nicht-codierenden Sequenzen mit einer PSSM. Da phylogenetisches Footprinting eine davon unabhängige Methode ist, sollte die Kombination der Profil-Information einer PSSM und der phylogenetischen Konserviertheit zwischen Mensch und Maus eine verbesserte Vorhersage von TFBSs erzielen. Bereits die einfachste Form der Kombination, d.h. eine Eingrenzung von PSSM-basierten Vorhersagen auf konservierte Bereiche, erzielte in den meisten untersuchten Fällen eine Verbesserung der Vorhersagequalität. Im Rahmen dieser Arbeit wurde ein HMM entworfen, das diese zwei unabhängigen Methoden zur Vorhersage von TFBSs in einer synergistischen Weise kombiniert. Das HMM wurde entsprechend der gewonnenen Erkenntnisse über die unterschiedliche Konserviertheit der TFBSs bestimmter TFs parametrisiert. Auf den untersuchten Testdatensätzen machte das HMM exaktere Vorhersagen als eine rein PSSM-basierte Suche nach TFBSs. In bestimmten Fällen wurde

im direkten Vergleich bei gleicher Sensitivität die Zahl falsch positiver Vorhersagen auf ein Viertel reduziert.

Die möglichst korrekte Vorhersage von TFBSs mit dieser Methode liefert einen Grundstein für die Konstruktion genregulatorischer Netzwerke und damit auch für ein besseres Verständnis der Transkriptionsregulation innerhalb einer Zelle.

Anhang A

Tabellen-Anhang

Tabelle A.1: Übersicht über die Ergebnisse der orthologen Sequenzbeschaffung und die Anzahl konservierter TFBSs für Datensatz I

	Anzahl	Prozent
TFBSs mit EMBL-Link	3508	100 %
- keine Mappierung	72	2.1 %
- nicht an ein Gen mappiert	53	1.5 %
- kein annotiertes orthologes Gen	224	6.4 %
WU-BLAST-Suche möglich	3159	100 %
- kein orthologer Treffer	481	15.2 %
- orthologe Sequenz gefunden	2678	84.8 %
Untersuchte TFBSs	2578	100 %
- Sequenz-konservierte TFBSs	1848	71.7 %
TFBSs mit zugehöriger PSSM	1890	100 %
- Muster-konservierte TFBSs	1314	69.5 %
- Sequenz-konservierte TFBSs	1366	72.3 %
- sequenz- und Muster-konservierte TFBSs	1104	58.4 %
- weder sequenz- noch Muster-konservierte TFBSs	314	16.6 %

Tabelle A.2: Sequenz-Konserviertheitsraten von 928 Mensch-TFBSs (Datensatz IV) für Speziesvergleiche zwischen Mensch und Maus, Ratte, Kuh oder Hund, aufgeschlüsselt nach Genen mit mindestens fünf TFBSs; Fortsetzung

Gen-Name	N^*	$C_{seq}^{\text{Maus}\dagger}$	$p_{seq}^{\text{Maus}\ddagger}$	$C_{seq}^{\text{Ratte}\dagger}$	$p_{seq}^{\text{Ratte}\ddagger}$	$C_{seq}^{\text{Kuh}\dagger}$	$p_{seq}^{\text{Kuh}\ddagger}$	$C_{seq}^{\text{Hund}\dagger}$	$p_{seq}^{\text{Hund}\ddagger}$
PGK1	7	100.0%	0.03287	100.0%	0.04234	57.1%	1	42.9%	1
IL12B	7	100.0%	0.03287	85.7%	0.59491	100.0%	0.03796	100.0%	0.02789
IL4	7	85.7%	0.48612	100.0%	0.04234	100.0%	0.03796	100.0%	0.02789
IL2RA	7	100.0%	0.03287	85.7%	0.59491	71.4%	1	71.4%	1
VIP	7	100.0%	0.03287	57.1%	1	85.7%	0.54536	100.0%	0.02789
CCL5	7	42.9%	1	57.1%	1	57.1%	1	71.4%	1
MITF	7	100.0%	0.03287	100.0%	0.04234	100.0%	0.03796	100.0%	0.02789
MAOB	6	50.0%	1	83.3%	1	83.3%	1	50.0%	1
HNF4A	6	66.7%	1	66.7%	1	33.3%	1	83.3%	1
F9	6	83.3%	1	83.3%	1	66.7%	1	66.7%	1
SERPINE1	6	100.0%	0.09762	50.0%	1	66.7%	1	83.3%	1
IFNG	6	33.3%	1	66.7%	1	66.7%	1	33.3%	1
IL5	6	50.0%	1	50.0%	1	66.7%	1	83.3%	1
CD2	6	33.3%	1	33.3%	1	50.0%	1	16.7%	1
DBH	6	83.3%	1	83.3%	1	66.7%	1	66.7%	1
CYP3A4	6	16.7%	1	50.0%	1	33.3%	1	83.3%	1
TERT	6	33.3%	1	33.3%	1	16.7%	1	16.7%	1
DCT	5	100.0%	0.28993	80.0%	1	80.0%	1	80.0%	1
PSEN1	5	80.0%	1	80.0%	1	40.0%	1	80.0%	1
PDGFB	5	80.0%	1	80.0%	1	40.0%	1	40.0%	1
CCL2	5	0.0%	1	40.0%	1	40.0%	1	40.0%	1
HAND1	5	60.0%	1	60.0%	1	0.0%	1	40.0%	1
FASLG	5	80.0%	1	100.0%	0.3474	80.0%	1	80.0%	1
AGT	5	40.0%	1	40.0%	1	60.0%	1	20.0%	1
IL10	5	40.0%	1	60.0%	1	60.0%	1	40.0%	1
PFKM	5	80.0%	1	80.0%	1	40.0%	1	80.0%	1
SPRR1B	5	60.0%	1	60.0%	1	80.0%	1	20.0%	1
PRL	5	100.0%	0.28993	100.0%	0.3474	100.0%	0.32135	100.0%	0.25787
PENK	5	100.0%	0.28993	100.0%	0.3474	80.0%	1	80.0%	1
HIST1H4A	5	0.0%	1	40.0%	1	80.0%	1	100.0%	0.25787
F8	5	60.0%	1	20.0%	1	20.0%	1	20.0%	1
ADA	5	80.0%	1	60.0%	1	60.0%	1	60.0%	1

*Anzahl der TFBSs pro TF.

†Die Hintergrund-Konserviertheitsrate C_{seq}^{bg} beträgt für Mensch-Maus-Vergleiche 33.67% (KS $= 65\%$), für Mensch-Ratte-Vergleiche 34.91% (KS $= 64\%$), für Mensch-Kuh-Vergleiche 34.37% (KS $= 76\%$) und für Mensch-Hund-Vergleiche 32.89% (KS $= 78\%$).

‡Die Wahrscheinlicheit p, die beobachtete (oder eine größere) Differenz zwischen C_{seq} und C_{seq}^{bg} zufällig zu erhalten, wurde mit Gleichung (3.7) berechnet. Die p-Werte wurden Bonferroni-korrigiert.

Anhang B

Abbildungs-Anhang

```
cow   : CCGGGGAGGGAGGGGGAAGAGAGGGGCGCGAGTCGCGC------------ @ 38/1171
dog   : GCGGAGGAGACGGGGGCAGAGCCGGGCGCGAGCCGCGCTTGAGGGGGAGG @ 50/1153
human : CCGGGGAGGGGACCGGGGAAGAGAGGGCCGAGAGGCGT------------ @ 38/1217
mouse : CCGGGGAGGGAACCCGGGAACACAAGCCGAAGCTGAGC------------ @ 38/1173
rat   : CCGGGGAGGGAACCCGGGAACACAAGCCGGAGCAACGC------------ @ 38/1165
      = 1         11        21        31        41
```

```
cow   : --GGAGGGGGGTCGGGGGGGGAAGGAGGAGAAAGAAGGGCCCAACTGTAG @ 86/1171
dog   : GGGGAGGGGGAGGGGGAGGGGGAGGAGGAGAAAGAAGGGCCCAACTGTAG @ 100/1153
human : ----GCGGCA----GGGGGGAGGGTAGGAGAAAGAAGGGCCCGACTGTAG @ 80/1217
mouse : ----GCGGGAGGGGGGGGGGGAGGAGGAGAAAGAAGGGCCCAACTGTAG @ 84/1173
rat   : ----GCGGGA---GGGGGAGGGAGGTGGAGAAAGAAGGGCCCAACTGTAG @ 81/1165
      = 51        61        71        81        91
```

```
                      R08483 (c-Jun)
cow   : GAGGGCAGCGGAGCATTACCTCATCCCGTGAGCCTCCGCGGGCCCAGAGA @ 136/1171
dog   : GAGGGCAGCGGAGCATTACCTCATCCCGTGAGCCTCCGCGGGGCCGGAGA @ 150/1153
human : GAGGGCAGCGGAGCATTACCTCATCCCGTGAGCCTCCGCGGGCCCAGAGA @ 130/1217
mouse : GAGCGCAGCGGAGCATTACCTCATCCCGTGAGCCTTCGCGGGCCCAGAGA @ 134/1173
rat   : GAGCGCAGCGGAGCATTACCTCATCCCGTGAGCCTTCGCGGGCCCAGAGA @ 131/1165
      = 101       111       121       131       141
```

```
                    R08501 (NF-jun)         R00955 (Sp1)
cow   : AGAATCTTCTAGGGTGGGGTATCCATGGCGACGGGTGGGCCCGCCCCCCT @ 186/1171
dog   : AGAATCTTCTAGGGTGGGGTCTCCATGGCGACGGGCGGGCCCGCCCCCTG @ 200/1153
human : AGAATCTTCTAGGGTGGAGTCTCCATGGTGACGGGCGGGCCCGCCCCCCT @ 180/1217
mouse : AGAATCTTCTAGGGTGAGGTCTCCATGGCGACGGGTGGGCCCGCCCCCCT @ 184/1173
rat   : AGAATCTTCTAGGGTGAGGTCTCCATGGCGACGGGTGGGCCCGCCCCCTT @ 181/1165
      = 151       161       171       181       191
```

```
                  R00956 (CTF)        R00957 (AP-1) R09147 (MEF-2B)
cow    : GAGAGCTACGCGAGCCAATGGGAAGGCCTTGGGGTGACATCATGGGCTAT @ 236/1171
dog    : GAGAGCGACGCGAGCCAATGGGAAGGCCTTGGGGTGACATCATGGGCTAT @ 250/1153
human  : GAGAGCGACGCGAGCCAATGGGAAGGCCTTGGGGTGACATCATGGGCTAT @ 230/1217
mouse  : GAGAACGACGCAAGCCAATGGGAAAGCCTCGGGGTGACATCATGGGCTAT @ 234/1173
rat    : GAGAGCGACGCAAGCCAATGGGAAGGCCTTGGGGTGACATCATGGGCTAT @ 231/1165
       = 201       211       221       231       241

   R09147 (MEF-2B)
cow    : TTTTAGGGGTTGACTGGTAGCAGATAAGTGTTGCGCTCCGGCTGGATAAG @ 286/1171
dog    : TTTTAGGGGTTGACTGGTAGCAGATAAGTGTTGAGCTCCGGCTGGATAAG @ 300/1153
human  : TTTTAGGGGTTGACTGGTAGCAGATAAGTGTTGAGCTCGGGCTGGATAAG @ 280/1217
mouse  : TTTTAGGGATTGACTGGTAGCAGATAAGTGTTGAGCTCAGGCTGGATAAG @ 284/1173
rat    : TTTTAGGGATTGACTGGTAGCAGATAAGTGTTGAGCTCAGGCTGGATAAG @ 281/1165
       = 251       261       271       281       291


cow    : GGTTCAGAGTTGCACAGAGTGTGGCTGAAGCTACGAGGCGGGAGTGGAGG @ 336/1171
dog    : GGCTCAGAGTTGCACTGAGTGTGGCCGAGGCGGCGAGGTGGGAGTGGAGG @ 350/1153
human  : GGCTCAGAGTTGCACTGAGTGTGGCTGAAGCAGCGAGGCGGGAGTGGAGG @ 330/1217
mouse  : GACTCAGAGTTGCACTGAGTGTGGCAGAGACAGCCTGGCAGGAGAGCGCT @ 334/1173
rat    : GACTCAGAGTTGCACTGAGTGTGGCAGAGACTGCCTAGCTGGAGAGCGCT @ 331/1165
       = 301       311       321       331       341


                          R09952 (SMAD)             M00792 (SMAD)
cow    : TGCGCGG----ACGCAGGCAGGCAGACAGGCACAGTCAGTCGGGA----- @ 377/1171
dog    : AGCGCGG----ACTCGGggaggcggcggCCGCGGGCAGGCGAGACGGCC @ 396/1153
human  : TGCGCGGAGTCAGGCAGACAGACAGACACAGCCAGCCAGCCAGGTCGGCA @ 380/1217
mouse  : CAGGCAG--ACAGACAGACAGACGGACGGACTTGGCCAACCCGGTCGGCC @ 382/1173
rat    : CAGG------CAGACAGACAGACGGACGGACTCGGCTAACCTGGTCGTCC @ 375/1165
       = 351       361       371       381       391


cow    : --------AGGACTGCAAATCC---TATTCTCCAATTT--------CTCT @ 408/1171
dog    : GGAGAGTCCGGGCTGCAAAATC---CGCTTTCCGTTTTGCATTTTCCTCG @ 443/1153
human  : GTATAGTCCGAACTGCAAATCTTATTTTCTTTTCACCT-------TCTCT @ 423/1217
mouse  : GCGGACTCCGGACTGTTCATCC--GTTTGTCTTCATTT-------TCTCA @ 423/1173
rat    : GCGGACTCCGGGCTGTTCATCT--GTTTGTCTTCATTT-------TCTGA @ 416/1165
       = 401       411       421       431       441


cow    : CCAACTGCCCGGAGCTAGCGTTTGTGGCTCCCGGCCTGGTGTTTTGGGGA @ 458/1171
dog    : CCGACTGGCCGGAGCTAGCGCCTGTGGCTCCCGGGCTGGTGTTTCGGGGA @ 493/1153
human  : CTAACTGCCCAGAGCTAGCGCCTGTGGCTCCCGGGCTGGTGTTTC-GGGA @ 472/1217
mouse  : CCAACTGCTTGGATCCAGCGCCCGCGGCTCCTGCACCGGTATTTTGGGGA @ 473/1173
rat    : CCAACTGCCTGGATCCAGCGCCCGCAGCTCCTGCACCGGTGTTTTGGGGA @ 466/1165
       = 451       461       471       481       491
```

```
cow   : GCGCCGGGAGAG-CCCCTTCTCCAGCCGCCCCCAGGCGGAGAGCCCCGCT @ 507/1171
dog   : GCGCCCGGAGAACCCCCTTCTCCAGCCGCCCCCGGGAGGAGAGCCCCGCC @ 543/1153
human : GTGTCCAGAGAG-CCTGGTCTCCAGCCGCCCCCGGGAGGAGAGCCCTGCT @ 521/1217
mouse : GCATTTGGAGAG-TCCCTTCTCCCGCCTTCCACGGAGAAGAAGCTCACAA @ 522/1173
rat   : GCACTTGGAGAG-TCCCTTCTCCCGCTTTTCCCGGAGAAGAAGCTCACAA @ 515/1165
      = 501       511       521       531       541


cow   : GCGCGGGCGCTGCTGACAGCGCGGAGAGCGG-------CTACTGTCCGCC @ 550/1171
dog   : GCCCGGGAGCCG-TGACAGCGCGGAGAGCGG-------CGACGGTCCGTC @ 585/1153
human : GCCCAGGCGCTGTTGACAGCGGCGGAAAGCAGCGGTACCCACGCGCCCGC @ 571/1217
mouse : GTCCGGGCGCTGCTGACAGCATCGAGAGCGG--------CTCCCGACCGC @ 564/1173
rat   : GTCCCGGCACTGCTGACAGCATCGAGAACGG--------CTCCGGACCGC @ 557/1165
      = 551       561       571       581       591


cow   : CGGGGAAGTCCGGAGAGCGGCTGCAGCGGCAAGAACTTTTCCCGGCCAGG @ 600/1171
dog   : CGCGCAAGTGCAGAGAGTGGCTGCGGCGGCTCAGAACTTTCCTGGCCGGG @ 635/1153
human : CGGGGGAAGTCGGCGAGCGGCTGCAGCAGCAAAGAACTTTCCCGGCTGGG @ 621/1217
mouse : GCGAGGAAATAGGCGAGCGGCTAC--CGGCCAGCAACTTTCCTGACCCAG @ 612/1173
rat   : GCGAGGAAATAGGCGAGCGGCTAC--CGGCCAGCAACTTTCCTGACCCAG @ 605/1165
      = 601       611       621       631       641


cow   : AAGACAGGAGACAAGTGG-----CCGCCGGGTCCCGGAACGAACTTTTGC @ 645/1171
dog   : AGCACAGGGGACGAGTGG-----C--CCGGGTCCCGGAGCGAACTTTTGC @ 678/1153
human : AGGACCGGAGACAAGTGGCAGAGT--------CCCGGAGCCAACTTTTGC @ 663/1217
mouse : AGGACCGGTAACAAGTGG--------------CCGGGAGCGAACTTTTGC @ 648/1173
rat   : AGGACCGGGAACAAGTGG--------------CCCGGAGCGAACTTTTGC @ 641/1165
      = 651       661       671       681       691


cow   : -AAACCTTGCCTCCGCC-CGAAGCTGCCGCGGCGGCGGAGAAG------- @ 686/1171
dog   : -AAAGCTCTGCCGCGCC-TGAGGCTGCGGCGGCGGCGGCGGCGGCGG @ 726/1153
human : -AAGCCTTTCCTGCGTC--TTAGGCTTCTCCACGGCGGTAAAG------- @ 703/1217
mouse : -AAATCTCTTCTGCGCCTTAAGGCTGCCACCGAGACTGTAAAG------- @ 690/1173
rat   : AAAACCTCTTCTGCGCC-TAAGGCTGCCACCGAGACCGTAAAG------- @ 683/1165
      = 701       711       721       731       741


cow   : ------------------AAAAAGAGGCGGCG------------------ @ 700/1171
dog   : CGGCGGCGGCCGGGACGAGAGAGGCGGCGGCG------------------ @ 758/1153
human : --------------------------ACCAGAAGGCGGCGGAGAGCCAC @ 726/1217
mouse : --------------------------AAAAG------------------ @ 695/1173
rat   : --------------------------AAAAG------------------ @ 688/1165
      = 751       761       771       781       791
```

```
cow   : --GAGAGAGGA----------GGACGTGCGCTCCGCTTCGCTGGCACCGG @ 738/1171
dog   : --GAGCGAGGAGACCGCGGAGGGACGTGCGCTCGGCGCCCCTCGCCCCGG @ 806/1153
human : GCAAGAGAAGA---------AGGACGTGCGCTCAGCTTCGCTCGCACCGG @ 767/1217
mouse : ---GGAGAAGA---------GGAACCTATACTCATACCAGTTCGCACAGG @ 733/1173
rat   : ---GGAGAAGA---------GGAACCTATACTCATACCAGCTCGCACAGG @ 726/1165
      = 801       811       821       831       841
```

```
cow   : TTGCTGAACTTGGGCGAGCGCGAGCCGCGACTGCCGGGCG-CCCCCTCCC @ 787/1171
dog   : TGGCTGAACTTGGGCGAGCGCGAGCCGCGGCTGCCCGGCG-CCCCCTCCC @ 855/1153
human : TTGTTGAACTTGGGCGAGCGCGAGCCGCGGCTGCCGGGCG-CCCCCTCCC @ 816/1217
mouse : CGGCTGAAGTTGGGCGAGCGCTAGCCGCGGCTGCCTAGCGTCCCCCTCCC @ 783/1173
rat   : CGGCTGAAGTTGGGCGAGTGCTAGCCGCGGCTGCCTAGCGTCCCCCTCCC @ 776/1165
      = 851       861       871       881       891
```

```
                                        M00469 (AP-2alpha)
cow   : CCTCGCAGCGGAGGAGGGGAC-AGTCTTCCGAGTCGGGGCGGCGGACACC @ 836/1171
dog   : CCTGTC--------------------------GCCGCGGGGGCGG---- @ 874/1153
human : CCTAGCAGCGGAGGAGGGGACAAGTCGTCGGAGTCCGGGCGGCCAAGACC @ 866/1217
mouse : CCTCACAGCGGAGGAGGGGAC-AGTTGTCGGAGGCCGGGCGGCAGAGCcc @ 832/1173
rat   : CCTCACAGCGGAGGAGGGGAC-AGTTGTCGGAGGCCGGGCGGCAGAGCcc @ 825/1165
      = 901       911       921       931       941
```

```
          R00959 (AP-2alphaA)
cow   : CGCCGCCGGCCGGCCACCGCGAAGTCCGCTCTCCCTCCTATCACCGGG-- @ 884/1171
dog   : ------------------------------------------------- @ 874/1153
human : CGCCGCCGGCCGGCCACTGCAGGGTCCGC----ACTGATCCGCTCCGCGG @ 912/1217
mouse : gatcgcgggct-TCCACCGAGAATTCCGTGACGACTGGTCAGCACCGC-C @ 880/1173
rat   : gatcgcgggct-TCCACCAAGAATTCCGTGACGACTGGTCTGCGCCGC-C @ 873/1165
      = 951       961       971       981       991
```

```
                                    R00958 (AP-2alphaA)
cow   : --------------AAGCGAGTT----CGTCTGCGGGCTCTGAGGAACC @ 915/1171
dog   : ----------------------------gcccgcgggCCCCCGGGAGCC @ 895/1153
human : GGAGAGCCGCTGCTCTGGGAAGTGAGTTCGCCTGCGGACTCCGAGGAACC @ 962/1217
mouse : GGAGAGCCGCTGTTGCTGGGACTG-----GTCTGCGGGCTCCAAGGAACC @ 925/1173
rat   : GGAGAACCTCTGTCGCTGGGGCTG-----gtccgcgggCTCCGAGGAACC @ 918/1165
      = 1001      1011      1021      1031      1041
```

```
                                    R00960 (AP-1)
cow   : GCTGCG-----------CTCGAGAGAGCTCCGTGAGTGACCGCGACTTT @ 953/1171
dog   : GCCGCGCCCCGAGAGCGCCCCGAGCGCGccccgtgggtgacCGCGACTTT @ 945/1153
human : GCTGCG-----------CACGAAGAGCGCTCAGTGAGTGACCGCGACTTT @ 1001/1217
mouse : GCTGCT-----------CCCCGAGAGCGCTCCGTGAGTGACCGCGACTTT @ 964/1173
rat   : GCTGCT-----------TCCCGAGAGCGCTCCGTGAGTGACCGCGACTTT @ 957/1165
      = 1051      1061      1071      1081      1091
```

```
                                        M00925 (AP-1)
cow   : T--CAAAGCCGGGCGGCGCGCGCG---AGCGGACAAGTAAGAGCGCGGGC @ 998/1171
dog   : TCCCCGAGGCGGGCAGCGCGCGCGCCAGCTGACCCAGGAGGAGCGCGG-- @ 993/1153
human : T--CAAAGCCGGGTAGCGCGCGCG---AGTCGACAAGTAAGAGTGCGGGA @ 1046/1217
mouse : T--CAAAGCTCGGCATCGCGCGGG---AGCCTACCAACGTGAGTGCTAGC @ 1009/1173
rat   : T--CAAAGCTCCGGATCGCGCGGG---AGCCAACCAACGTGAGTGCAAGC @ 1002/1165
      = 1101      1111      1121      1131      1141


cow   : GGCGCC---TTAACCCTGCGATCCCCCGGACTGAGCTGGTGAGGAGGACG @ 1045/1171
dog   : ----------------------CGGCGGCCGGAGCCGGTGAGCGGGGCG @ 1020/1153
human : GGCATCTTAATTAACCCTGCGCTCCCTGGAGCGAGCTGGTGAGGAGGGCG @ 1096/1217
mouse : GGAGTC----TTAACCCTGCGCTCCCTGGAGCGAACTGGGGAGGAGGGCT @ 1055/1173
rat   : GGTGTC----TTAACCCTGCGCTCCCTGGAGCGAACTGGGGAGGAGGGCG @ 1048/1165
      = 1151      1161      1171      1181      1191


cow   : CCGGGGGCGGGGGCTGCCAGCCGGCGGGGGCGCGCGCTCTTCCAGAAACT @ 1095/1171
dog   : CGGGGGGAGGA------CCGCCGGCGGGAGCGCGCGCCCCTCCAGAAACT @ 1064/1153
human : CAGCGGGGACG-----ACAGCCAGCGGGTGCGTGCGCTCTTAGAGAAACT @ 1141/1217
mouse : CAGGGGGAAGC-----ACTGCCGTCTGGAGCGCACGCTCCTAAACAAACT @ 1100/1173
rat   : CAGGGGGGAGC-----ACTGCGGTCTGGAGCGCACGCTCCTAAACAAACT @ 1093/1165
      = 1201      1211      1221      1231      1241
```

Abbildung B.1: Multiples Alignment des 5'-UTR-Bereichs des *JUN*-Gens von Kuh, Hund, Mensch, Maus und Ratte. Position 1 des Alignments entspricht der TSS. Rot hervorgehoben sind neun für den Menschen annotierte TFBSs. Nukleotide der anderen Spezies, die zu diesen TFBSs aligniert und konserviert sind, sind fett hervorgehoben. Für andere Spezies annotierte TFBSs sind in blau hervorgehoben. Wenn die Sequenzen, die zu den Mensch-TFBSs aligniert sind, einen PI-Wert unter dem entsprechenden KS aufweisen, sind diese in der jeweiligen Spezies mit kursiven Kleinbuchstaben dargestellt. In der Umgebung der nicht konservierten TFBS gefundene PSSM-Treffer für den gleichen TF sind in grün dargestellt. Nukleotide der anderen Spezies, die zu diesen vorhergesagten TFBSs aligniert und konserviert sind, sind auch fett hervorgehoben.

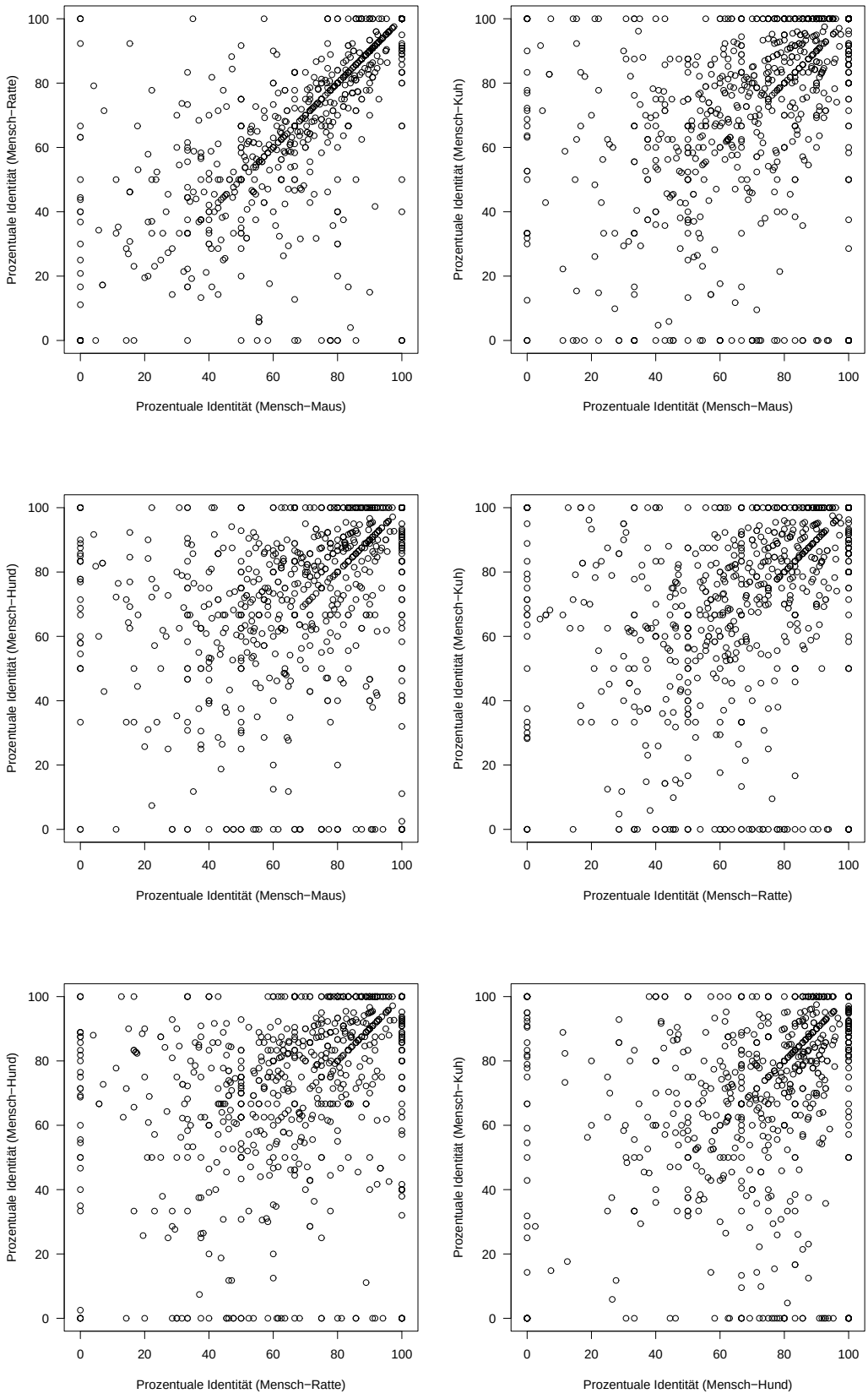

Abbildung B.2: Korrelation der PI-Werte der TFBSs aus Datensatz IV für verschiedene Speziesvergleiche. Die Korrelation ist für Mensch-Maus- und Mensch-Ratte-Vergleiche mit einem Pearson-Korrelationskoeffizienten von $R = 0.74$ am höchsten.

Literaturverzeichnis

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–410. 3.2.3.1

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402. 3.2.3.1, 3.2.3.2

Batzoglou S (2005) The many faces of sequence alignment. *Brief Bioinform* **6**, 6–22. 3.2.3

Bedell J, Korf I, Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040–1041. 3.3.3

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent W, et al. (2004) Ultraconserved elements in the human genome. *Science* **304**, 1321–1325. 4.2.2

Berezikov E, Guryev V, Plasterk R, Cuppen E (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res* **14**, 170–178. 3.2.3.6

Bergman C, Pfeiffer B, Rincon-Limas D, Hoskins R, Gnirke A, et al. (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol* **3**, research0086. 2.2.3

Berman B, Pfeiffer B, Laverty T, Salzberg S, Rubin G, et al. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5**, R61. 2.2.3

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. *Nucleic Acids Res* **34**, 556–561. 3.1.4

Bland J, Altman D (1995) Multiple significance tests: the Bonferroni method. *BMJ* **310**, 170. 4.2.1

Bray N, Dubchak I, Pachter L (2003) AVID: A global alignment program. *Genome Res* **13**, 97–10. 3.2.3.4

Brudno M, Do C, Cooper G, Kim M, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**, 721–731. 3.2.3.8

Brudno M, Morgenstern B (2002) Fast and sensitive alignment of large genomic sequences. *Proc IEEE Comput Soc Bioinform Conf* **1**, 138–147. 3.2.3.8

Buck M, Lieb J (2005) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *2004* **83**, 349–360. 2.2

Bulyk M (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol* **5**, 201. 2.2.3

Cavener D (1987) Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res* **15**, 1353–1361. 2.2.1

Chen L (1999) Combinatorial gene regulation by eukaryotic transcription factors. *Curr Opin Struct Biol* **9**, 48–55. 2.1.3

Chen YQ, Sengchanthalangsy LL, Hackett A, Ghosh G (2000) NF-kappaB p65 (RelA) homodimer uses distinct mechanisms to recognize DNA targets. *Structure* **8**, 419–428. 2.1.3

Chiaromonte F, Yap V, Miller W (2002) Scoring pairwise genomic sequence alignments. *Pac Symp Biocomput* , 115–126. 3.2.3.2

Choi S, Bush E, Lahn B (2005) Different classes of tissue-specific genes show different levels of noncoding conservation. *Genomics* **87**, 433–436. 4.2.2

Choo K, Vissel B, Nagy A, Earle E, Kalitsis P (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res* **19**, 1179–1182. 2.2.1

Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, et al. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res* **34**, 10–15. 3.1.1

Cooper G, Sidow A (2003) Genomic regulatory regions: insights from comparative sequence analysis. *Curr Opin Genet Dev* **13**, 604–610. 2.2.3

Cormen TH, Stein C, Rivest RL, Leiserson CE (2001) Introduction to Algorithms. The MIT Press. 2.2.4

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, et al. (2004) The Ensembl automatic gene annotation system. *Genome Res* **14**, 942–950. 3.1.4

Day W, McMorris F (1992a) Consensus sequences based on plurality rule. *Bull Math Biol* **54**, 1057–1068. 2.2.1

Day W, McMorris F (1992b) Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res* **20**, 1093–1099. 2.2.1

Dermitzakis E, Clark A (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19**, 1114–1121. 2.2.3, 4.2

Dermitzakis E, Reymond A, Lyle R, Scamuffa N, Ucla C, et al. (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**, 578–582. 2.2.3

Duret L, Bucher P (1997) Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* **7**, 399–406. 2.2.3

Elnitski L, Hardison R, Li J, Yang S, Kolbe D, et al. (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res* **13**, 64–72. 2.2.3

Emberly E, Rajewsky N, Siggia E (2003) Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics* **4**, 57. 2.2.3

Feng D, Doolittle R (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25**, 351–360. 3.2.3.5

Frazer K, Elnitski L, Church D, Dubchak I, Hardison R (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res* **13**, 1–12. 2.2.3

Frith MC, Li MC, Weng Z (2003) Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* **31**, 3666–3668. 4.4

Gotoh O (1982) An improved algorithm for matching biological sequences. *J Mol Biol* **162**, 705–708. 3.2.3

Hannenhalli S, Levy S (2002) Predicting transcription factor synergism. *Nucleic Acids Res* **30**, 4278–4284. 4.4

Hardison R, Oeltjen J, Miller W (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7**, 959–966. 2.2.3

Hardison R, Xu J, Jackson J, Mansberger J, Selifonova O, et al. (1993) Comparative analysis of the locus control region of the rabbit beta-like gene cluster: HS3 increases transient expression of an embryonic epsilon-globin gene. *Nucleic Acids Res* **21**, 1265–1272. 4.2.1

Harris M, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258–D261. 3.3.7

Hermfisse U, Schafer D, Netzker R, Brand K (1996) The aldolase A promoter in proliferating rat thymocytes is regulated by a cluster of SP1 sites and a distal modulator. *Biochem Biophys Res Commun* **225**, 997–1005. 4.2.1

Huang X, Miller W (1991) A time-efficient, linear-space local similarity algorithm. *Adv Appl Math* **12**, 337–357. 3.2.3.9

Iwama H, Gojobori T (2004) Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *Proc Natl Acad Sci USA* **101**, 17156–17161. 4.2.2

Jareborg N, Birney E, Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* **9**, 815–824. 4.1

John B, Enright AJ, Aravin A, Tuschl T, Sander C, et al. (2004) Human MicroRNA targets. *PLoS Biol* **2**. 3.1.4

Jordan I, Rogozin I, Glazko G, Koonin E (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19**, 68–72. 4.1

Jurka J (2000) Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**, 418–420. 3.3.3

Kel A, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis O, et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **31**, 3576–3579. 2.2.2

Kel A, Kel-Margoulis O, Babenko V, Wingender E (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol* **288**, 353–376. 3.2.2

Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res,* **30**, 332–334. 3.1.2

Kellis M, Patterson N, Endrizzi M, Birren B, Lander E (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254. 2.2.3

Kent W, Dubchak I, Pachter L (2002) BLAT–the BLAST-like alignment tool. *Genome Res* **12**, 656–664. 3.2.3.3

Lee S, Kohane I, Kasif S (2005) Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics* **6**, 168. 4.2.2

Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, et al. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J Biol* **2**, 1. 2.2.3, 4.2

Levy S, Hannenhalli S (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* **13**, 510–514. 2.2.3, 4.2

Lewin B (2002) Molekularbiologie der Gene. Heidelberg, Deutschland: Spektrum Akademischer Verlag. 2.1.2

Li J (2003) AtomEye: an efficient atomistic configuration viewer. *Modelling Simul Mater Sci Eng* **11**, 173–177. 2.1

Liu Y, Liu X, Wei L, Altman R, Batzoglou S (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res* **14**, 451–458. 2.2.3, 4.1, 4.2

Loots G, Locksley R, Blankespoor C, Wang Z, Miller W, et al. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140. 2.2.3

Ludwig M, Bergman C, Patel N, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567. 4.2, 4.2.1

Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. *Genome Biol* **1**. 2.1.3

Marinescu VD, Kohane IS, Riva A (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics* **6**, 79–79. 2.2.5

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374–378. 3.1.1

Mirny LA, Gelfand MS (2002) Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res* **30**, 1704–1711. 4.2.3

Morgenstern B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218. 3.2.3.7

Morgenstern B, Dress A, Werner T (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci USA* **93**, 12098–12103. 3.2.3.7

Morgenstern B, Frech K, Dress A, Werner T (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14**, 290–294. 3.2.3.7

Moses A, Chiang D, Kellis M, Lander E, Eisen M (2003) Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol* **3**, 19. 2.2.3, 4.2.3

Müller CW (2001) Transcription factors: global and detailed views. *Curr Opin Struct Biol* **11**, 26–32. 2.1.3

Needleman S, Wunsch C (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443–458. 3.2.3, 3.2.3

Notredame C (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* **3**, 131–144. 3.2.3

Notredame C, Higgins D, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205–217. 3.2.3.10

Notredame C, Holm L, Higgins D (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics* **14**, 407–422. 3.2.3.10

Pollard D, Bergman C, Stoye J, Celniker S, Eisen M (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**, 6. 4.2

Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, et al. (2004) The Ensembl analysis pipeline. *Genome Res* **14**, 934–941. 3.1.4

Prakash A, Tompa M (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat Biotechnol* **23**, 1249–1256. 2.2.3, 4.1

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical Recipes in C: The Art of Scientific Computing. New York, NY, USA: Cambridge University Press. 4.2.1, 4.2.3

Prestridge DS, Stormo G (1993) SIGNAL SCAN 3.0: new database and program features. *Comput Appl Biosci* **9**, 113–115. 2.2.2

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**, 4878–4884. 2.2.2, 3.2.2

Rippe RA, Brenner DA, Tugores A (2001) Techniques to measure nucleic acid-protein binding and specificity. Nuclear extract preparations, DNase I footprinting, and mobility shift assays. *Methods Mol Biol* **160**, 459–479. 2.2

Robertson A, Bilenky M, Lin K, He A, Yuen W, et al. (2006) cisRED: A database system for genome scale computational discovery of regulatory elements. *Nucleic Acids Res* **34**, D68–D73. 3.1.3, 4.2

Rosenberg M (2005) Evolutionary distance estimation and fidelity of pair wise sequence alignment. *BMC Bioinformatics* **6**, 102. 4.2

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425. 3.2.3.5

Sandelin A, Bailey P, Bruce S, Engstrom P, Klos J, et al. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99. 4.2.2

Sandelin A, Wasserman WW (2005) Prediction of nuclear hormone receptor response elements. *Mol Endocrinol* **19**, 595–606. 2.2.5

Schwabe JW (1997) The role of water in protein-DNA interactions. *Curr Opin Struct Biol* **7**, 126–134. 2.1.3

Schwartz S, Kent W, Smit A, Zhang Z, Baertsch R, et al. (2003) Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103–107. 3.2.3.2

Sinha S, Schroeder M, Unnerstall U, Gaul U, Siggia E (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics* **5**, 129. 2.2.3

Smith T, Waterman M (1981) Comparison of biosequences. *Adv Appl Math* **2**, 482–489. 3.2.3

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**, 1–10. 3.2.3.7

Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, et al. (2004) The Ensembl core software libraries. *Genome Res* **14**, 929–933. 3.1.4

Stormo G (2000) DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23. 2.2.2

Stormo G, Schneider T, Gold L (1982a) Characterization of translational initiation sites in E. coli. *Nucleic Acids Res* **10**, 2971–2996. 2.2.2

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982b) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **10**, 2997–3011. 2.2.2

Tagle D, Koop B, Goodman M, Slightom J, Hess D, et al. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203**, 439–455. 2.2.3

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793. 4.5

Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680. 3.2.3.5

Ureta-Vidal A, Ettwiller L, Birney E (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4**, 251–262. 2.2.3

Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Info Theory* **IT-13**, 260–269. 2.2.5, 3.4.1.3

Wasserman W, Palumbo M, Thompson W, Fickett J, Lawrence C (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* **26**, 225–228. 2.2.3, 4.2.1

Waterman M, Arratia R, Galas D (1984) Pattern recognition in several sequences: consensus and alignment. *Bull Math Biol* **46**, 515–527. 2.2.1

Waterman M, Eggert M (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol* **197**, 723–728. 3.2.3.9

Wingender E (1997) Classification of eukaryotic transcription factors. *Mol Biol (Mosk)* **31**, 584–600. 2.1.3

Woolfe A, Goodson M, Goode D, Snell P, McEwen G, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**, e7. 4.2.2

Xie X, Lu J, Kulbokas E, Golub T, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345. 2.2.3

Xu D, Liu HJ, Wang YF (2005) BSS-HMM3s: an improved HMM method for identifying transcription factor binding sites. *DNA Seq* **16**, 403–411. 2.2.5

LEBENSLAUF

Name	Tilman Sauer
Geburtsdatum	26.06.1976
Geburtsort	Gifhorn
Staatsangehörigkeit	deutsch

Schulbildung

1982-1986	Alfred-Teves-Grundschule in Gifhorn
1986-1988	Christoph-Kolumbus-Orientierungsstufe in Gifhorn
1988-1995	Otto-Hahn-Gymnasium in Gifhorn

Zivildienst

1995-1996	Zivildienstleistender beim Notfunkdienst Gifhorn e.V.

Studium

1996-1998	Grundstudium im Diplomstudiengang Chemie an der Technischen Universität Braunschweig
1998-2002	Hauptstudium im Diplomstudiengang Chemie an der Technischen Universität Braunschweig
September 1999 bis März 2000	Auslandsstudium an der University of Wales, Swansea
Juni bis Juli 2001	Diplomprüfungen in Anorganischer Chemie, Organischer Chemie, Physikalischer Chemie und Biochemie & Biotechnologie
September 2001 bis März 2002	Diplomarbeit im Bereich Sicherheitstechnik der BASF AG, Ludwigshafen

Berufstätigkeit

März bis August 1999	Studentische Hilfskraft in der Abteilung „Genregulation und Differenzierung“ der Gesellschaft für biotechnologische Forschung (GbF), Braunschweig
März bis November 2000	Studentische Hilfskraft an der Technischen Universität Braunschweig im Projekt „Vernetztes Studium Chemie“ des Bundesministeriums für Bildung und Forschung (BMBF)
Mai bis Juli 2002	Wissenschaftliche Hilfskraft am Insitut für Hochfrequenztechnik der Technischen Universität Braunschweig
Juli bis Dezember 2002	Wissenschaftlicher Angestellter der Abteilung Bioinformatik der GbF, Braunschweig
Seit 1.1.2003	Wissenschaftlicher Angestellter der Abteilung Bioinformatik des Universitätsklinikums der Georg-August-Universität Göttingen