Modellierung regulatorischer Netzwerke von Säugetieren und Einsatz von Methoden zur strukturellen Analyse und Identifikation von Kernkomponenten

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

vorgelegt von

Björn Goemann

aus Einbeck

Göttingen 2011

D 7

Referent:	Prof. Dr. Edgar Wingender
Korreferent:	Prof. Dr. Stephan Waack
Tag der mündlichen Prüfung:	20.04.2011

Danksagung

An dieser Stelle möchte ich mich bei allen herzlichst bedanken, die mich während dieser Arbeit begleitet und unterstützt haben. Als erstes danke ich Prof. Dr. Edgar Wingender, der Erstgutachter dieser Arbeit ist und sie überhaupt erst ermöglicht hat. Er war ein wertvoller Ansprechpartner, durch dessen Anregungen und umfangreichen wissenschaftlichen Erfahrungsschatz diese Arbeit sehr profitiert hat. Ein großer Dank gilt Dr. Anatolij Potapov für die Betreuung der Arbeit und dafür, dass er sich in vielen Stunden die Zeit genommen hat, mit mir zu diskutieren und mir Hilfestellung zu geben. Bei Professor Dr. Stephan Waack möchte ich mich dafür bedanken, Zweitgutachter dieser Arbeit zu sein. Tilman, Martin, Isolde und Martin danke ich vielmals für das kritische Lesen der Arbeit und ihre Korrekturvorschläge. Ich danke außerdem der gesamten Abteilung für Bioinformatik für die freundliche Atmosphäre und Hilfsbereitschaft, die mir den Arbeitsalltag enorm erleichtert haben. Meinem Büromitbewohner Martin Haubrock danke ich für seine Geduld, die vielen Diskussionen und das entspannte miteinander während dieser Zeit. Ebenso danke ich Torsten Schöps, der die Rechenmaschinen wacker am laufen hält. Carmen Modrok und Doris Waldmann danke ich für ihre Hilfe in administrativen Dingen. Schließlich möchte ich mich ganz besonders bei Heiko, Martin, Andrea, Annika, Alex, Tilman, Torsten, Sven sowie allen anderen, namentlich hier nicht genannten mir nahestehenden Menschen für ihre Freundschaft bedanken und dafür, dass sie für mich dagewesen sind. Mein tiefer Dank gilt meiner Familie, die mich immer auf meinem Weg bestärkt hat.

Inhaltsverzeichnis

Abkürzungsverzeichnis

C. elegans	*Caenorhabditis elegans*
E. coli	*Escherichia coli*
S. cerevisiae	*Saccharomyces cerevisiae*
ADP	Adenosindiphosphat
ATP	Adenosintriphosphat
d.h.	das heißt
DNA	Desoxyribonukleinsäure
i.d.R.	in der Regel
Muster	topologisches Muster
o.g.	oben genannte
PDI	*pairwise disconnectivity index*
RNA	Ribonukleinsäure
TF	Transkriptionsfaktor
TLR4	Toll-like Rezeptor 4
TLR7	Toll-like Rezeptor 7
u.a.	unter anderem
usw.	und so weiter
Vgl.	Vergleiche
z.B.	zum Beispiel

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

Trotz rasanter Forschritte in der Biotechnologie in den letzten Jahren sind die intrazellulären Abläufe in höheren Eukaryoten immer noch nicht vollständig verstanden. Das Sammeln adäquater Datenmengen für aussagekräftige Analysen ist eine schwierige Aufgabe, weil diverse Faktoren, wie der konkrete Zelltyp und die Zugehörigkeit zu einem bestimmten Gewebe, einen individuellen Mix an aktiven Genen in den Zellen multizellulärer Lebewesen ergeben. Dabei sind die lebenswichtigen intrazellulären Kreisläufe der Signaltransduktion, Genexpression und des Metabolismus bereits an sich durch die enorme Vielfalt an Regulationsmöglichkeiten in Mehrzellern wesentlich komplexer, als die einfacherer Organismen.

Signaltransduktion subsumiert die Vorgänge, die den Eintritt äußerer Reize in Form von Signalen in die Zelle ermöglichen und für ihre Weiterleitung sorgen. Abhängig vom initialen Stimulus werden unterschiedliche, teils eng miteinander verflochtene Signalkaskaden getriggert, wodurch Änderungen in der Expression der Gene und des Stoffwechsels als zelluläre Reaktion ausgelöst werden. Die Steuerung der Genexpression ist von zentraler Bedeutung, da die Genprodukte Bestandteile aller regulatorischen Programme sind. Sie erfolgt über die Aktivierung oder Inhibierung von Transkriptionsfaktoren, die entscheiden, ob und in welcher Intensität Gene exprimiert werden. Für die chemische Umwandlung von Stoffen im Rahmen des Metabolismus sind Genexpression und Signaltransduktion insofern relevant, weil die zur Katalyse metabolischer Prozesse unersetzbaren Enzyme genomkodiert sind und ihre Aktivität durch äußere Reize beeinflusst wird. Im Metabolismus lenken Stoffwechselwege u.a. die Gewinnung und Bereitstellung der für den Ablauf der meisten biologischen Reaktionen notwendigen Energie.

Die Betrachtung derartiger Mechanismen in einer Netzwerkperspektive hat das Verständnis über ihre grundlegenden Funktionsweisen stark gefördert. Solche Netzwerke können auf sehr spezifische Aspekte regulatorischer Kreisläufe, beispielsweise im Kontext einer Krankheit, fokussiert sein oder wesentliche Vorgänge in einer Zelle lediglich umreißen. Für diesen Zweck sind graphentheoretische Ansätze allgemein als mächtige Werkzeuge bekannt, weil sie eine einfache Modellierung erlauben, mit der sich eine maßgeschneiderte Abstraktion erzeugen lässt. Aus der topologischen Analyse der generierten Graphen können nützliche Hinweise über die Zusammenhänge in den abgebildeten Systemen

gewonnen werden, die vor allem wegen der oftmals nur sehr aufwendigeren oder auch gar nicht möglichen Realisierung der entsprechenden Laborexperimente von unschätzbarem Wert sind. In den letzten Jahren haben sich graphentheoretische Studien deshalb zunehmender Popularität erfreut.

Für großen Anklang sorgten die Arbeiten von *Alon* [1] und *Barabasi* [2], die signifikante strukturelle Unterschiede zwischen den von ihnen untersuchten biologischen und zufällig erzeugten Netzwerken aufzeigen konnten. Mit der Skalenfreiheit, enormen Robustheit gegenüber Störungen und der Überrepräsentation bestimmter topologischer Muster, die sogenannten Motive, wurden daraus mehrere, für regulatorische Systeme als generisch empfohlene Attribute abgeleitet. Die Netzwerke, für deren Architektur solche Eigenschaften berichtet wurden, reichen von metabolischen Netzwerken in Archaea, Bakterien und Eukaryoten [3, 4, 5, 6, 7, 8] über Protein-Protein Netzwerke aus Hefe und Fliege [3, 5, 9, 10, 11] bis zu einigen Genexpressions- und Signalnetzwerken [9, 10, 12, 13, 14, 15]. Demnach scheinen sich funktional sehr unterschiedliche Systeme fundamentale Merkmale zu teilen, obwohl die dahinterstehenden Konzepte zum Teil kontrovers diskutiert werden [16, 17, 18].

Das aktuelle Wissen über die Topologie biologischer Netzwerke basiert jedoch fast ausschließlich auf der Analyse von Prokaryoten und einzelligen Eukaryoten. Über die Prinzipien, nach denen die komplexeren regulatorischen Kreisläufe in höheren Eukaryoten aufgebaut sind, ist bisher kaum etwas bekannt. Ob für ihre Architektur die gleichen oder doch gänzlich andere Gesetzmäßigkeiten gelten, muss sich erst noch zeigen. Im Rahmen der vorliegenden Arbeit soll dieser Fragestellung, bezogen auf die Steuerung der Genexpression durch Transkriptionsfaktoren, der intrazellulären Übertragung von Signalen und dem Metabolismus nachgegangen werden. Eine der Herausforderungen dabei ist, experimentell nachgewiesene Aktivitäten einzelner Gene/Proteine zu aussagekräftigen Netzwerken über diese Programme zusammenzuführen. Aufgrund der Verfügbarkeit an Daten ist dies durch eine differenzierte Modellierung nach Spezies, Gewebe, Zelltyp, usw. nicht möglich. Veröffentlichungen über regulatorische Prozesse in höheren Eukaryoten beziehen sich allen voran auf die Spezies Mensch, Maus und Ratte, so dass es einen Weg zu finden gilt, der die Daten dieser drei Spezies miteinander verknüpft.

Eine weitere Herausforderung dieser Arbeit ergibt sich daraus, dass klassischer Weise einzelne Gene oder Proteine im Vordergrund von Laboruntersuchungen stehen, deren Funktion zum Beispiel im Kontext einer bestimmten Krankheit

untersucht werden soll. Die enorme Anzahl an potenziellen Kandidaten macht aber die Auswahl eines bestimmten Gens, dass experimentell überprüft werden soll, zu einem mühsamen Unterfangen. Für Experimentalisten ist es deswegen außerordentlich hilfreich, eine Vorselektion durch die Analyse theoretischer Modelle zu erhalten. Da viele der theoretischen Abhandlungen hauptsächlich die Eigenschaften ganzer Systeme hervorheben, ist die Anzahl an Methoden zur Bestimmung von Schlüsselknoten in Graphen gering, so dass ein großer Bedarf an Metriken besteht, die diese Lücke zu schließen vermögen.

2 Zellübergreifende mammalische Netzwerke

Nach derzeitigem Kenntnisstand besitzt das menschliche Genom zwischen 20.000 und 30.000 Gene, in denen die notwendigen Informationen zur Synthese von Proteinen hinterlegt sind [19, 20]. Wie bei den meisten Eukaryoten können durch alternatives Splicing aus einem einzigen Gen mehrere Proteine hergestellt werden, so dass die Größenordnung an genkodierten Proteinen im Menschen bei etlichen hundertausend liegt. Sowohl die Expression eines Gens, als auch die Aktivität eines Proteins hängt im Allgemeinen nicht nur vom Zustand einer Zelle ab, sondern auch vom Zelltyp, Gewebe, Organ sowie diversen inneren und äußeren Einflüssen. Unter welchen Bedingungen ein bestimmtes Gen oder Protein benötigt wird, ist entsprechend aufwendig im Labor nachzuweisen, weshalb die Menge an verfügbaren Daten nur sehr spärlich ist.

Dieselbe Komplexität trifft grundsätzlich auch auf die regulatorischen Programme in anderen Säugetieren zu. Dafür können insbesondere bei Mäusen und Ratten Laboruntersuchungen durchgeführt werden, die beim Menschen nur mühsam oder überhaupt nicht realisierbar sind. Von Vorteil daran ist, dass sich die meisten Erkenntnisse auf den Menschen übertragen lassen. Verglichen mit Mäusen und Ratten ist beispielsweise die Anzahl an Genen im menschlichen Genom in etwa gleich groß [21]. Hinzu kommt der enorme Anteil an orthologen proteinkodierenden Genen, d.h. solchen, die aus einem gemeinsamen Vorläufergen stammen und daher für ähnliche Aminosäuresequenzen kodieren: Bei Mensch und Maus liegt er bei ca. 80% [20], bei Mensch und Ratte sogar bei knapp 90% [21]. Sehr wahrscheinlich besteht eine hohe Übereinstimmung hinsichtlich der Funktion der Genprodukte in Mensch, Maus und Ratte.

Dieser Umstand lässt sich zur Modellierung der essentiellen Mechanismen - Transkription, Signaltransduktion und Metabolismus - in einer zell- und speziesübergreifenden Sichtweise nutzen, die für eine einzige Spezies unter Berücksichtigung von Zelltyp, Gewebe, usw. mangels Daten wenig erkenntnisreich wäre. Im Vordergrund dieser abstrakteren Betrachtungsebene steht damit der prinzipielle Ablauf dieser Programme, ohne auf die Eigenheiten in einer Spezies einzugehen. Als Grundlage dienen die orthologen Gene in Säugetieren (Mensch, Maus und Ratte) bzw. die experimentell beobachteten Interaktionen zwischen ihnen oder ihren kodierten Proteinen. Getrennt nach Transkription, Signaltransduktion und Metabolismus werden die beobachteten Zusammenhänge in den einzelnen Spezies in gerichtete Graphen mit folgender Knoten- und

Kantensemantik übertragen: *i*) Orthologe Gene bzw. die Proteine solcher Gene werden stets zu einem speziesübergreifenden Repräsentanten als Knoten zusammengefasst. Ein Gen (Protein) kommt daher genau einmal in einem Graphen vor. *ii*) Die Kanten bündeln die Menge an verfügbaren Interaktionen zwischen den Genen (Proteinen) in den einzelnen Spezies und illustrieren somit den Kausalzusammenhang, d.h. die Wechselwirkung zweier Gene (Proteine) unabhängig von den exakten chemischen Reaktionen. Eine Kante kann daher auf vielen, experimentell nachgewiesenen Interaktionen in jeder der drei Spezies beruhen, aber auch bislang nur für eine einzige Spezies bekannt sein.

Transkription

Transkriptionsfaktoren sind in jedem Lebewesen essentiell für die Synthese von RNA aus einem Gen (Transkription), aus der im weiteren Verlauf der Genexpression ein Protein hergestellt werden kann. Unterschieden wird bei Eukaryoten zwischen *allgemeinen* und *spezifischen* Transkriptionsfaktoren. Während allgemeine Transkriptionsfaktoren hauptsächlich für den Ablauf der Transkription verantwortlich sind, steuern spezifische Transkriptionsfaktoren deren Effizienz, so dass sie letztlich die Durchführung zellspezifischer Programme wie Zellwachstum oder -differenzierung kontrollieren. Meistens reguliert ein und derselbe Transkriptionsfaktor zahlreiche Gene, wobei er die Expression jedes Gens fördern oder blockieren kann. Spezifische Transkriptionsfaktoren wirken jedoch in der Regel nicht einzeln, sondern als individuell zusammengesetzter Verbund je Gen im Sinne einer regulatorischen Einheit. Dies bringt vor allem eine beträchtliche kombinatorische Vielfalt mit sich, die die gezielte Regulation jedes einzelnen Gens in Abhängigkeit vom jeweiligen zellulären Kontext ermöglicht.

Die Expression eines spezifischen Transkriptionsfaktors wird auch in der gleichen Art und Weise von mehreren Transkriptionsfaktoren gesteuert, die selber wiederum von anderen reguliert werden. Die gegenseitige Regulation von Genen, die für Transkriptionsfaktoren kodieren, stellt daher einen eigenen Kreislauf dar, der zumindest in Teilen vor der Expression jedes Gens aktiv ist und deshalb für die Genregulation an sich von enormer Bedeutung ist. Das mammalische **Transkriptionsnetzwerk** beschreibt diesen Kreislauf auf Grundlage der experimentell verifizierten Inhalte der Datenbanken TRANSFAC® (Version 11.3) und TRANSPATH® (Version 8.3). Während in TRANSFAC® Informationen über eukaryotische *cis*-regulatorische DNA-Bindestellen für Transkriptionsfaktoren gespeichert sind [22], enthält TRANSPATH® Informationen über mammalische Signaltransduktionsprozesse [23].

Im abstrahierten Transkriptionsnetzwerk sind die 279 Knoten jeweils eine speziesübergreifende Abstraktion orthologer Gene in den Spezies Mensch, Maus und Ratte. Beispielsweise ist der in Abbildung 1 skizzierte c-*fos* Knoten die Superposition der c-*fos* Gene in Mensch, Maus und Ratte, die dort den c-Fos Transkriptionsfaktor kodieren. Die 657 Kanten beschreiben Interaktionen der Transkriptionsfaktoren an unterschiedlichen DNA-Bindestellen. Eine Kante mit gleichem Start- und Zielgen entspricht der Autoregulation eines Gens, d.h. ein Transkriptionsfaktor beschleunigt oder bremst die Expression seines eigenen Gens. 63 Gene sind im Transkriptionsnetzwerk autoregulatorisch. Eine Kante zwischen zwei verschiedenen Genen („*c-fos reguliert c-jun*“) bedeutet dagegen, dass der Transkriptionsfaktor vom Gen am Beginn der Kante (*c-fos*) mit dem Gen am Ende der Kante (*c-jun*) interagiert und so die Expression seines kodierten Transkriptionsfaktors beeinflusst.

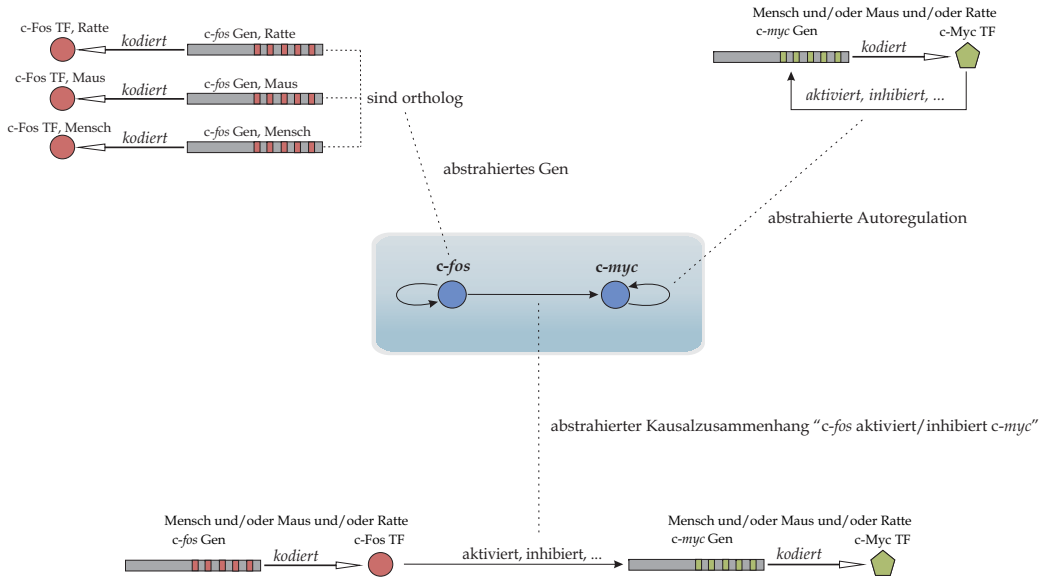


Abbildung 1: Modellierungsschema des Transkriptionsnetzwerks.

Signaltransduktion

Etliche, durch Transkriptionsfaktoren regulierte Gene kodieren für Proteine, die Funktionen in der Signaltransduktion besitzen. Signaltransduktion ist ein essentieller Mechanismus, der Zellen eine Reaktion auf Umwelteinflüsse (Licht, Hitze) oder Reize (Hormone, Neurotransmitter, Bluthochdruck) ermöglicht. Signaltransduktionsprozesse werden durch solch einen Stimulus in Gang gesetzt und wandeln ihn in eine zellspezifische Antwort um. Dazu wird das von außen auf die Zelle einwirkende Signal durch eine Serie von biochemischen Reaktionen

zum Zielort (z.B. Zellkern) transportiert. Die ausgelöste zelluläre Reaktion kann beispielsweise in Änderungen der Genexpression oder des Stoffwechsels bestehen, deren Folge komplexe zelluläre Programme wie eine Immunreaktion oder Apoptose einleiten können.

Die Verarbeitung von Signalen beginnt i.d.R. mit ihrer Aufnahme durch Rezeptoren. Intrazelluläre Rezeptoren binden bereits in der Zelle befindliche Signale[1], Membranrezeptoren hingegen extrazelluläre[2]. Danach startet die eigentliche Signaltransduktion, die die Weiterleitung von Signalen zu den intrazellulären Zielen regelt. Typischerweise läuft die Signalweiterleitung über mehrere hintereinandergeschaltete Protein-Protein Interaktionen als sogenannte Signalkaskade ab, die in jedem einzelnen Schritt eine Verstärkung bzw. Verbreitung des Signals (Amplifikation) bewirken kann oder einen insgesamt sehr spezifischen Effekt (z.B. auf einen Transkriptionsfaktor) auslöst.

An den Protein-Protein Interaktionen sind vor allem Proteinkinasen beteiligt. Proteinkinasen sind Enzyme, die durch Phosphorylierung Proteine modifizieren und damit ihren Aktivitätszustand verändern oder ihnen den Zugang in den Zellkern erlauben. Ein und dasselbe Protein kann aufgrund von Modifikationen (Phosphorylierung, Ubiquitinierung, usw.) in mehreren Formen vorliegen, so dass die wirkliche Anzahl an Signalmolekülen enorm ist und sich die Menge an ablaufenden Prozessen dementsprechend potenziert. Die exakte Abbildung der durch Enzyme katalysierten chemischen Reaktionen endet entsprechend schnell in großen, schlecht handhabbaren Modellen. In wissenschaftlichen Veröffentlichungen hat sich deshalb ein simplifiziertes Schema etabliert, dass lediglich die Weiterleitung der Signale zwischen Proteinen, die direkt an der Prozessierung der Signale mitwirken, skizziert [24]. Im Vordergrund steht somit der Signalfluss anstatt der genauen chemischen Annotation der Vorgänge inklusive aller beteiligten Moleküle[3].

Diese vereinfachte Darstellung macht sich auch die TRANSPATH® Datenbank zu Nutze, aus deren Inhalten das mammalische **Signaltransduktionsnetzwerk**

[1]Sekundäre Botenstoffe wie Calcium oder einige Liganden (z.B. Steroidhormone), die die Zellwand durch Diffusion passieren können.

[2]Membranrezeptoren sind an der Zelloberfläche und haben eine extrazelluläre Bindedomäne, an die ein Ligand andocken kann; die dadurch herbeigeführte Konformationsänderung des Rezeptors bewirkt auf der Innenseite der Zellmembran die Aktivierung der entsprechenden Signalkaskade.

[3]Beispielsweise läßt sich die Phosphorylierung eines Proteins `P` durch eine Kinase `K` unter Verwendung von ATP von $\mathtt{P} + \mathtt{ATP} + \mathtt{K} - - > \mathtt{P}(\mathtt{phosphoryliert}) + \mathtt{ADP} + \mathtt{K}$ auf *K (de-)aktiviert P* reduzieren.

erstellt wurde. Das Netzwerk beschreibt, wie die in Signalen gebundenen Informationen zwischen genkodierten Signalproteinen bis zu den Transkriptionsfaktoren in mammalischen Zellen weitergeleitet werden (Abbildung 2). Die 1571 Knoten sind wiederum Abstraktionen von Signalproteinen, deren kodierende Gene jeweils Orthologe in Mensch, Maus und Ratte sind. Genkodierte Signalproteine können u.a. Rezeptoren, Liganden, Adaptoren, Enzyme und Transkriptionsfaktoren sein, nicht-genkodierte Signalmoleküle, wie z.B. sekundäre Botenstoffe, werden nicht explizit als eigenständige Knoten modelliert.

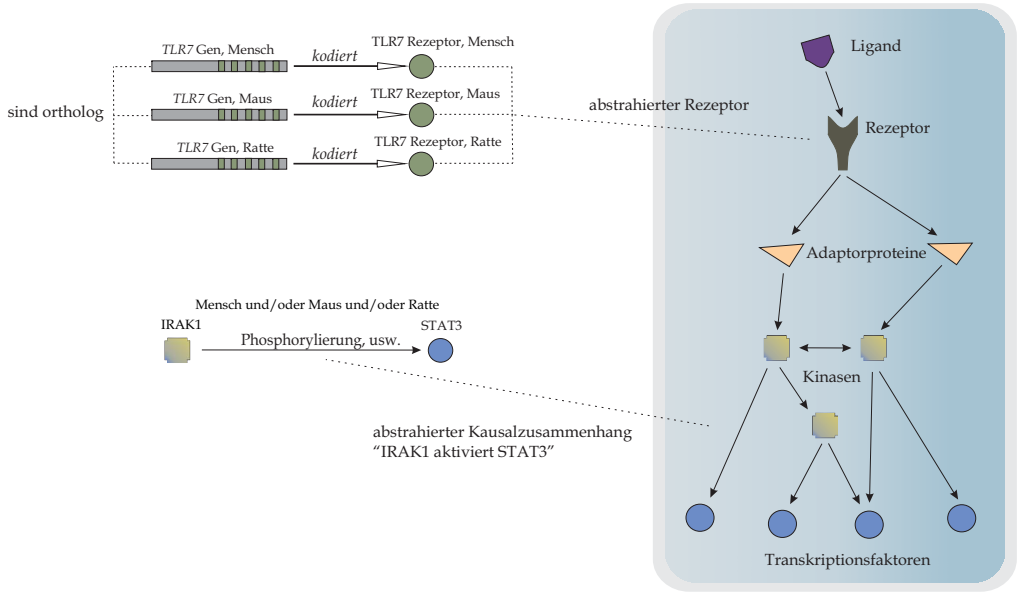


Abbildung 2: Modellierungsschema des Signaltransduktionsnetzwerks.

Kanten verkörpern den Kausalzusammenhang zwischen den Signalproteinen, der hier besagt, dass es eine oder mehrere molekulare Reaktionen in mindestens einer der drei Spezies gibt, in denen das Protein am Anfang einer Kante auf das am Ende einer Kante wirkt[4]. Stimmen Start- und Zielprotein einer Kante überein, reguliert sich ein Protein *autokatalytisch*, d.h. es ist gleichzeitig Katalysator und Produkt einer Reaktion. 53 der insgesamt 3425 Kanten sind autokatalytisch.

Metabolismus

Die Katalyse chemischer Prozesse durch Enzyme findet nicht nur bei der Signalverarbeitung statt, sondern ist der Schlüsselmechanismus im Stoffwechsel

[4]Sind noch weitere genkodierte Signalproteine an den Reaktionen beteiligt, werden entsprechend von diesen aus Kanten zu den Produkten der Reaktionen gezogen.

von Lebewesen überhaupt. Enzyme senken den Energiebedarf, der zum Initiieren einer Reaktion notwendig ist, erheblich und beschleunigen die gesamte Reaktion um ein Vielfaches, ohne dabei selbst verbraucht oder verändert zu werden. Viele Reaktionen würden ohne diese Katalysatoren entweder gar nicht oder nur extrem langsam ablaufen.

Im Metabolismus katalysieren Enzyme die chemische Umwandlung von Stoffen, wobei grundsätzlich zwischen dem Abbau von Stoffen (Katabolismus) und dem Aufbau körpereigener Stoffe (Anabolismus) unterschieden wird. Die Transformation von Stoffen erfolgt im Metabolismus nicht willkürlich, sondern wohl strukturiert in Stoffwechselwegen. Ein Stoffwechselweg besteht aus einer Serie aufeinanderfolgender chemischer Reaktionen, die insgesamt eine definierte Funktion ausüben, z.B. die Aufspaltung eines bestimmten Stoffes (Glukose, Fettsäuren). Charakteristisch für Stoffwechselwege ist die Verkettung der durch Enzyme katalysierten Reaktionen: Das Produkt einer Reaktion ist stets zugleich Substrat für die nächste Reaktion. Die Endprodukte eines Stoffwechselweges können anschließend sowohl in weiteren metabolischen Stoffwechselwegen eingesetzt werden oder sind in Form von Energieträgern, Signalmolekülen (z.B. sekundäre Botenstoffe), usw. unverzichtbare Bausteine für Signaltransduktion, Genregulation und andere regulatorische Programme. Im Gegenzug wirken sich Änderungen in diesen Programmen auf die Ausführung der Stoffwechselvorgänge aus. Beispielsweise können aufgrund eines hormonellen Impulses Signalkaskaden getriggert werden, in denen Proteinkinasen durch Phosphorylierung metabolische Enzyme *an- oder ausschalten*. Zusätzlich lassen sich Stoffwechselwege durch die Anpassung der Expressionsstärke von Genen, die für metabolische Enzyme kodieren, kontrollieren.

Die Datenbank *Kyoto Encyclopedia of Genes and Genomes* (KEGG) enthält mittlerweile ein sehr umfangreiches Repertoire über die Teilnahme von genetisch kodierten Enzymen an metabolischen Prozessen [25]. Basierend auf diesen Daten wurde das **metabolische Netzwerk** in Zusammenarbeit mit Michael Ante erstellt, welches die wechselseitigen Abhängigkeiten zwischen Enzymen, die metabolische Reaktionen in mammalischen Zellen katalysieren, in einer Sicht veranschaulicht, die analog zum Signaltransduktionsnetz die genkodierten Komponenten (hier: metabolische Enzyme) ins Zentrum rückt. Knoten sind daher die kodierenden Gene dieser Enzyme und jeweils Superpositionen der Orthologe in den Spezies Mensch, Maus und Ratte. Kanten beruhen auf den Abhängigkeiten der metabolischen Reaktionen in den Stoffwechselwegen

(Abbildung 3): Das kodierte Enzym vom Gen am Anfang einer Kante katalysiert einen Stoff, der als Substrat in eine Reaktion eingeht, die vom kodierten Enzym des Gens am Ende der Kante katalysiert wird. 94 Kanten mit gleichem Start- und Zielgen stehen hier für reversible Reaktionen, bei der ein Enzym seinen katalysierten Stoff selber konsumiert. Insgesamt besteht das Netzwerk aus 1793 Knoten und 5538 Kanten.

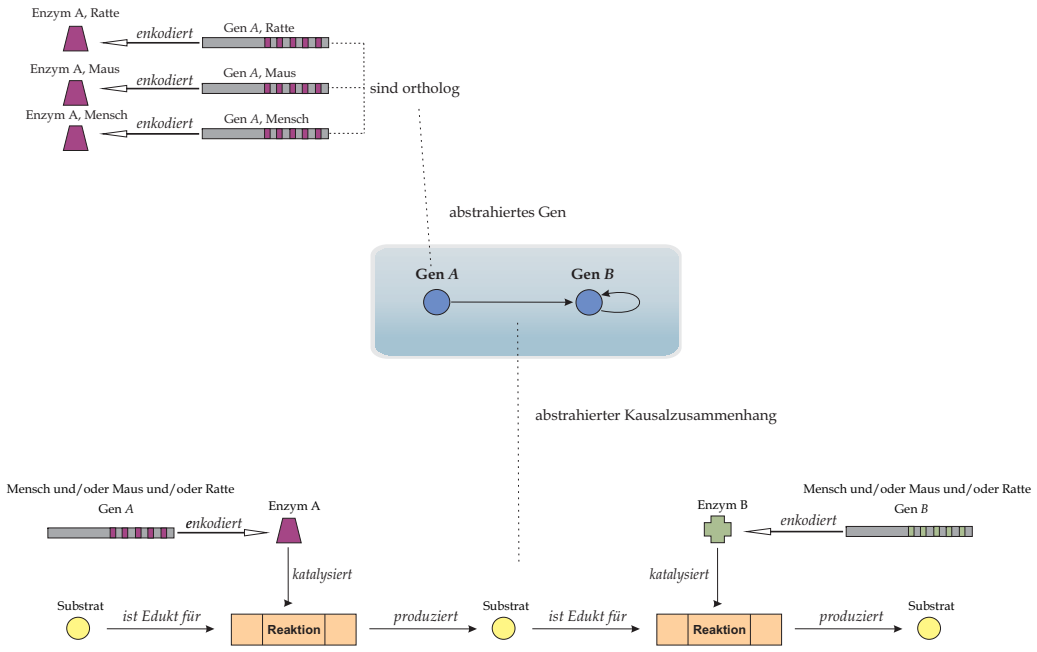


Abbildung 3: Modellierungsschema des metabolischen Netzwerks.

3 Die Verteilung des *degrees*

Als eine der bedeutendsten Erkenntnisse hinsichtlich des strukturellen Aufbaus biologischer Netzwerke wurde in den letzten Jahren hervorgehoben, dass die durch Knoten abstrahierten Gene/Moleküle nicht willkürlich miteinander verknüpft sind [2]. Im Gegensatz zu Zufallsnetzwerken, in denen die meisten Knoten in etwa die gleiche Anzahl von Kanten (englisch: *degree*)[5] haben, gibt es in biologischen Systemen eine große Menge von Knoten mit kleinem *degree*, während wenige andere, die sogenannten Hubs, einen enorm hohen *degree* haben. Für die Verteilung des *degrees*, der die Wahrscheinlichkeit des Auftretens von *degree* k, $P(k)$, angibt, wurde daher statt der für Zufallsnetzwerke charakteristischen Poissonverteilung ein Potenzgesetz (*power-law*) der Form

$$P(k) \sim k^{-\gamma}/\zeta(\gamma) \tag{3.1}$$

gemessen, bei dem γ für eine Konstante und $\zeta(\gamma)$ für die Riemann-Zeta Funktion mit $\zeta(\gamma) = \sum_{k=1}^{\infty} k^{-\gamma}$ steht [2]. Großen Anklang fand diese Beobachtung vor allem deshalb, weil ein *power-law* verteilter *degree*[6] eine Vorraussetzung für Skalenfreiheit ist. Skalenfreie Netzwerke sind in allen Teilen stets gleich aufgebaut, d.h. die Struktur eines Subsystems entspricht der des Gesamten, so dass die Systemparameter unabhängig von der Größe des Netzwerks sind [26]. Mit der Skalenfreiheit werden zudem ein hierarchisch modularer Aufbau sowie eine enorme Robustheit gegenüber Störungen in Verbindung gebracht. Insgesamt festigten die mit skalenfreien Systemen assoziierten Eigenschaften maßgeblich die Vorstellung einer nach bestimmten Prinzipien aufgebauten, generischen Architektur biologischer Netzwerke [27, 28]. Ihr Ursprung wird im Wachstum eines Netzwerks nach dem Mechanismus des *preferential attachment* gesehen, bei dem die Interaktionspartner neuer Gene mit einer Wahrscheinlichkeit proportional zu ihrem *degree* gewählt werden und somit ein *power-law* verteilter *degree* generiert wird [2].

Neuere Studien weisen jedoch darauf hin, dass den Verteilungen der *degrees* vieler Netzwerke oftmals voreilig ein Verlauf nach dem *power-law* Modell unterstellt wurde [29, 30]. Als Hauptursache wird die üblicherweise zur Approximation

[5] Im folgenden wird der englische Begriff *degree* anstelle von Grad verwendet, da im deutschen auch die Begriffe *in-degree* für den Eingangsgrad eines Knotens und *out-degree* für den Ausgangsgrad gebräuchlich sind. Der Begriff *inout-degree* steht für die Summe von Ein- und Ausgangsgrad eines Knotens.

[6] Im Allgemeinen wird mit einem *power-law* verteilten *degree* bzw. dem Begriff *power-law* Modell eine Verteilung nach Formel 3.1 verbunden.

verwendete lineare Regression im doppelt logarithmierten Plot angesehen, die wegen etlicher Mängel zu einer fehlerhaften Anpassung führen kann [30, 31]. Des Weiteren wird meist ausschließlich auf eine weitestgehende Ähnlichkeit zum *power-law* Modell geachtet, ohne ein möglicherweise besseres Fitting durch ein anderes Modell zu testen. Wenig überraschend ist deshalb vielleicht auch, dass erneute Analysen mit der verlässlicheren Maximum-Likelihood-Methode für viele Netzwerke ihre ursprüngliche Klassifizierung als skalenfreies System aufgrund einer schlechten Übereinstimmung mit der *power-law* Verteilung zurückweisen [17, 29, 32]. Stattdessen konnten wesentlich bessere Näherungen entweder durch eine Exponentialverteilung[7] oder ein *power-law* mit exponentiellem Ausläufer[8] erreicht werden [29].

Für die mammalischen Netzwerke stellt sich demnach ebenfalls die Frage, ob aus den *degree*-Verteilungen Skalenfreiheit abgeleitet werden kann, sie eher einen exponentiellen Verlauf haben oder sogar Zufallsnetzwerke sind. Letzteres kann für die Verteilungen der *in*-, *out*- und *inout-degrees* sofort ausgeschlossen werden: Es liegt kein poissonverteilter Verlauf vor, da die relative Häufigkeit des Auftretens eines bestimmten *degrees* mit seiner Höhe sinkt (Abbildung 4). Demnach interagieren Transkriptionsfaktoren, Signalproteine und metabolische Enzyme meistens nur mit wenigen anderen direkt (*inout-degree*), hochaktive Knoten, die *inout-degree* Hubs, sind die Ausnahme. Bemerkenswert ist für diese Hubknoten in allen drei Netzwerken die Antikorrelation von *in*- und *out-degree*, wodurch Gene/Proteine, die im gleichen Maßstab, wie sie reguliert werden, auch selber direkt auf andere Gene/Proteine wirken, praktisch fehlen[9]. Die hinter den Hubknoten stehenden Gene/Proteine haben daher entweder eine spezifische Wirkung, die auf verschiedenen Wegen veranlasst werden kann oder müssen gezielt aktiviert bzw. inhibiert werden, um einen koordinierten, umfangreichen Effekt auszulösen.

Die mit der Maximum-Likelihood Methode berechneten Approximationen der *degree*-Verteilungen durch o.g. Modelle[10] zeigen für die Netzwerke der Transkription und des Metabolismus, dass das *power-law* Modell keine akzeptable Näherung bietet, diese aber umso mehr durch die exponentiellen Modelle gegeben ist. Während sich im metabolischen Netzwerk für den *in*- und *inout-degree* das beste Fitting durch ein *power-law* mit exponentiellem Ausläufer erzielen lässt,

[7] $P(k) \sim e^{-\lambda k} / \sum_{k=1}^{\infty} e^{-\lambda k}$ mit λ als Konstante.
[8] $P(k) \sim k^{-\alpha} e^{-\beta k} / \sum_{k=1}^{\infty} k^{-\alpha} e^{-\beta k}$ mit α und β als Konstanten.
[9] Vgl. hierzu auch Abbildung A.1 in Anhang A.
[10] Vgl. hierzu den Abschnitt *Methods* in [33].

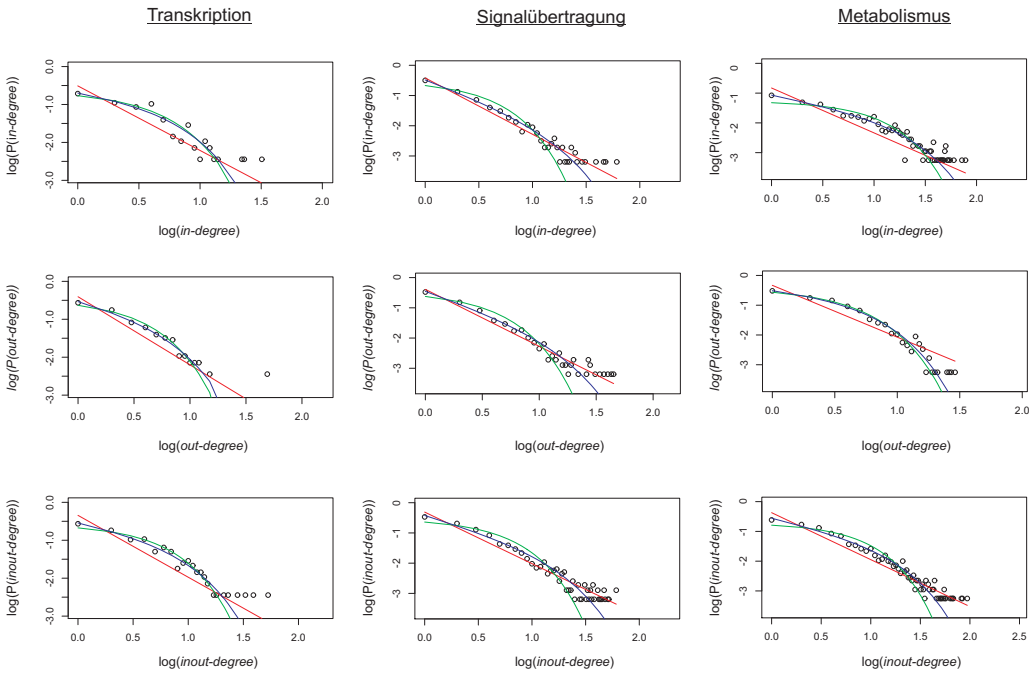


Abbildung 4: Die Verteilungen der *in*-, *out*- und *inout-degrees* in den mammalischen Netzwerken. In logarithmierten Skalen sind die *degrees* gegen die beobachteten relativen Häufigkeiten geplottet. Zusätzlich eingezeichnet sind die Approximationen der *degree*-Verteilungen durch das *power-law* Modell (rote Linien), die Exponentialverteilung (grüne Kurven) und das *power-law* mit exponentiellem Ausläufer (blaue Kurven).

fällt vor allem beim Transkriptionsnetzwerk die äußerst gute Übereinstimmung mit der Exponentialverteilung auf. Letztere ist zwar bereits für den Verlauf des *in-degrees* der Transkriptionsnetzwerke von Hefe [16] und *E. coli* [34, 35] beobachtet worden, aber nicht für den ihrer *out-degrees*, was in der weniger komplexen Steuerung der Genexpression durch Transkriptionsfaktoren im Vergleich zu höheren Eukaryoten begründet ist[11]. Ebenfalls keine gute Anpassung ergibt die Approximation durch die Exponentialverteilung für die *degree*-Verteilungen des mammalischen Signalnetzwerks. Stattdessen ist hier das *power-law* Modell eher geeignet, welches in Kombination mit einem exponentiellem Ausläufer die besten Näherungen liefert.

Unterstützung finden die aus Abbildung 4 gesammelten Eindrücke in den Resultaten eines Likelihood-Ratio Tests, der die Güte der Approximationen der verschiedenen Modelle anhand ihrer Maximum-Likelihood Werte bewertet [33]. Wird beispielsweise für das Transkriptionsnetzwerk das *power-law* Modell

[11] Beispielsweise liegen in *E. coli* mehrere Gene direkt hintereinander auf der DNA als Operon, welches durch einen bestimmten Satz an Transkriptionsfaktoren reguliert und an einem Stück abgelesen wird.

gegenüber dem exponentiellen eindeutig verworfen, verhält es sich erwartungsgemäß genau umgekehrt beim Signalnetzwerk. Ebenso eindeutig bevorzugt der Test allerdings auch das *power-law* Modell mit exponentiellem Ausläufer für die *degree*-Verteilungen des Signalnetzwerks.

Letzten Endes lässt sich damit keine der beobachteten *degree*-Verteilungen durch das *power-law* Modell nach Formel 3.1 approximieren, wodurch auch die Skalenfreiheit als fundamentale Eigenschaft der mammalischen Netzwerke verworfen werden kann. Von den drei getesteten Modellen liefert stattdessen das *power-law* mit exponentiellem Ausläufer stets die beste Anpassung, die Variabilität der *degree*-Verteilungen macht es jedoch schwierig, diese oder gar eine andere Verteilung als die universell gültige für biologische Netzwerke zu identifizieren [36].

4 Ein Verfahren zur Detektion von Schlüsselknoten

Zentralitätsmaße aus der Graphentheorie sind überaus hilfreich, wenn es darum geht, die Rolle einzelner Knoten zu bemessen. Sie können danach getrennt werden, ob sie Knoten in ihrem lokalen Umfeld bewerten oder im globalen Kontext des jeweiligen Netzwerks sehen. Ansätze mit lokalem Fokus, wie der *degree*, analysieren Knoten in einem engen Radius um sie herum. Sie heben damit die Bedeutung eines Knotens für einen kleinen Bereich eines Netzwerks hervor, wodurch ihre Aussagekraft allerdings stets auf diesen begrenzt ist. Globale Parameter sind dagegen frei von der Beschränkung auf die unmittelbare Nachbarschaft eines Knotens. Sie bewerten einen Knoten im Gesamtkontext eines Netzwerks und berücksichtigen insofern auch seinen indirekten Einfluss, der sich über viele Knoten hinweg erstrecken kann. Für die Analyse biologischer Systeme ist das von Vorteil, weil sich die Bedeutung eines Gens oder Proteins zu großen Teilen aus der indirekten Wirkungsweise ergibt[12].

Mehrere Studien deuten an, dass von den Zentralitätsmaßen mit globalem Fokus einzig die *betweenness centrality* Methode vielversprechende Ergebnisse zur Bestimmung von Schlüsselknoten in regulatorischen Netzwerken liefert [37, 38, 39, 40]. *Betweenness centrality* bemisst den Stellenwert eines Knotens an der Anzahl der kürzesten Wege zwischen je zwei unterschiedlichen Knoten, in denen er enthalten ist [41, 42, 43]. Große *betweenness centrality* Werte unterstreichen somit die Bedeutung eines Knotens als Mediator für kürzestmögliche Verbindungen. Allerdings verbergen sich in dem Verfahren zwei Limitationen, die seine Eignung zur Analyse regulatorischer Netzwerke schmälern. Erstens werden kürzeste Wege als die wichtigsten angesehen, was eine starke und zugleich irreführende Vereinfachung ist. In biologischen Systemen wird die Bedeutung eines Pfades nicht so sehr durch seine Länge, sondern durch die Effizienz und verwendete Semantik der dahinterstehenden Reaktionen bestimmt. Daher können längere Wege wesentlich schneller und effizienter als kürzere sein[13]. Zweitens kann *betweenness centrality* nur auf Knoten angewendet werden, die zwischen zwei anderen liegen. Periphere Knoten, die keine ein- oder ausgehenden Kanten haben, werden damit automatisch ignoriert. Dadurch sind beispielsweise

[12] Typische Fragestellungen sind *„Welche Transkriptionsfaktoren werden durch einen bestimmten Impuls aktiviert oder geblockt?“* oder *„Was verursacht der Knockout von Gen X?“*

[13] Abgesehen davon ist es gerade in genregulatorischen Netzwerken oftmals schwer, überhaupt eine Pfadlänge zu definieren, da die Kanten überlicherweise mehrere Reaktionen zusammenfassen.

in Signalnetzwerken extrazelluläre Liganden und Zielgene von einer Analyse ausgeschlossen.

Als neues Zentralitätsmaß versucht der *pairwise disconnectivity index* (PDI) Unzulänglichkeiten der existierenden Ansätze, allen voran die der *betweenness centrality*, zu vermeiden [44]. Das Verfahren erinnert an Methoden wie *vertex-connectivity* oder *edge-connectivity*, die den Grad der Vernetzung eines Graphen wiedergeben [45]. Die Grundidee beim PDI ist die globale Bedeutung eines Knotens anhand seiner **topologischen Signifikanz** zu beurteilen, die ausdrückt, wie essentiell er für die Vernetzung eines Graphen ist. Sie lässt sich quantifizieren, indem der Knoten aus seinem Graphen eliminiert und der sich daraus ergebene Effekt auf die Anzahl der verbundenen, geordneten Knotenpaare gemessen wird. In einem gerichteten Graphen heißen zwei Knoten i und j, $i \neq j$, verbundenes, geordnetes Knotenpaar $\{i, j\}$, wenn es mindestens einen Pfad von Knoten i zu Knoten j gibt[14]. Je mehr solcher Knoten-zu-Knoten Verbindungen vollständig gekappt wurden (j kann nicht mehr von i erreicht werden), umso größer ist die topologische Signifikanz des gelöschten Knotens. Für einen Knoten v ist der PDI gegeben durch

$$PDI(v) = 1 - \frac{N'}{N} \tag{4.1}$$

und beziffert, inwiefern sich die Anzahl N an verbundenen, geordneten Knotenpaaren im Graphen G von der Anzahl N' im Graphen G', der sich von G durch das Fehlen von Knoten v und aller seiner ein- und ausgehenden Kanten unterscheidet, ändert. Der maximale PDI von 1 bedeutet, dass kein Knoten mehr mit einem anderen verbunden ist[15] und somit Knoten v substanziell für die Vernetzung von G ist. Demgegenüber besagt ein PDI von 0, dass alle Knoten-zu-Knoten Verbindungen auch weiterhin bestehen.

Die Methode hat den Vorteil, dass sie sich leicht einsetzen lässt, ohne vorher irgendwelche Simplifizierungen an einem Netzwerk vornehmen zu müssen. Es werden nur die Kausalzusammenhänge zwischen den Knoten hervorgehoben, weshalb keine Abhängigkeit vom gewählten Abstraktionsgrad, der von Kante zu Kante variieren kann, besteht. Der Ansatz ist vergleichbar mit der Vorgehensweise im Labor, bei der ein Gen durch die Zugabe eines passenden Inhibitors ausgeschaltet und der darauffolgende Effekt gemessen wird. Der PDI

[14] In einem gerichteten Graphen gilt zudem $\{i, j\} \neq \{j, i\}$.

[15] Alternative Interpretation: 100% der ursprünglich existierenden Knoten-zu-Knoten Verbindungen sind unterbrochen.

kann benutzt werden, um ebenso gezielt die Rolle eines Gens zu evaluieren oder um ohne jegliche Fokussierung ein Netzwerk nach den interessantesten Genen zu filtern, deren Funktionen anschließend im Labor weiterverfolgt werden können. Das Grundprinzip des Verfahrens ist zudem problemlos auf andere Netzwerkelemente (Kanten oder Gruppen von Knoten/Kanten) übertragbar und kann auf die gleiche Art und Weise wie in Formel 4.1 bemessen werden (Kapitel 6 und [44, 46]). Eine Java Applikation sowie ein Webserver [47], welche die verschiedenen Varianten des PDI implementieren, stehen unter [48] und [49] zur Verfügung.

Die Leistungsfähigkeit des PDI zur Detektion von Schlüsselknoten ist umfassend für die Transkriptionsnetzwerke des Bakteriums *E. coli* [50] und der Hefe *S. cerevisiae* [1], sowie für das neuronale Netzwerk des Nematoden *C. elegans* [51] und das mammalische Toll-like Rezeptor 4 (TLR4) Signalnetzwerk [52, 53] untersucht worden [44]. Für alle vier Netzwerke zeigte der Vergleich mit der *betweenness centrality* Methode, dass beide Verfahren zwar für einige Knoten zu einer ähnlichen Einschätzung gelangen, was ihre Rolle in den Netzwerken angeht, *betweenness centrality* aber etliche Schlüsselknoten übersieht und ihre Bedeutung somit unterschätzt. Beispielsweise sind im TLR4 Netzwerk die meisten Proteine mit hohen PDI-Werten dafür bekannt, bei Entwicklung und Erhalt der notwendigen Lebensfunktionen entscheidend mitzuwirken, was in Knockoutexperimenten ihrer kodierenden Gene nachgewiesen wurde. Im Netzwerk des Bakteriums *E. coli* haben die Gene *crp*, *fnr* und *fis* einen *betweenness centrality* Wert von 0, weisen aber mit die höchsten PDI-Werte auf. Allein *crp* regelt die Expression von mehr als 100 Genen [54] und ist zusammen mit *fnr* und *fis* einer der globalen Transkriptionsregulatoren in *E. coli*, die die Expression von 51 % aller Gene kontrollieren [55]. Der PDI ist damit insgesamt eine leicht interpretierbare Alternative zu den bestehenden Methoden, dessen Beurteilungsmaßstab - der Beitrag zum Erhalt der Konnektivität - sich als geeignet für die Identifikation zentraler Knoten in regulatorischen Netzwerken erwiesen hat.

5 Störanfälligkeit und die Rolle der Hubknoten

Aus den PDI-Werten der Knoten in einem Netzwerk lässt sich auch erkennen, ob ein Netzwerk allgemein sensibel auf Störungen reagiert. Für die Netzwerke obiger Studie [44] hat sich beispielsweise gezeigt, dass nur eine Hand voll an Knoten einen großen Einfluss auf die Knoten-zu-Knoten Verbindungen ausübt, während die überwiegende Mehrheit aufgrund ihrer niedrigen PDI-Werte kaum relevant ist. Demzufolge sind Knoten-zu-Knoten Verbindungen in diesen Netzwerken äußerst robust gegenüber den meisten Perturbationen, reagieren aber empfindlich auf das gezielte Entfernen weniger Knoten. Robustheit ist ein existenzielles Merkmal biologischer Systeme, die ihnen erlaubt, ihre Funktionen in einer dynamischen Umgebung mit ungewissen Schwankungen fortzusetzen und sich weiterzuentwickeln. Sie gewährleistet eine große Ausfallsicherheit gegen in- und externe Störungen und beruht auf einer Architektur, die verschiedene Organisationsprinzipien vereinigt und so ein ausgewogenes Maß an Redundanz und Flexibilität bietet [56]. Frühere Simulationen haben in diesem Zusammenhang bereits auf eine generelle Widerstandsfähigkeit biologischer Systeme hingewiesen, die mit einer hohen Empfindlichkeit gegenüber gezielten Attacken auf die Hubknoten einhergeht [4, 26]. Fortan wurde den Hubknoten die zentrale Rolle in biologischen Systemen zugesprochen, deren Robustheit auf ihre vermeintliche Skalenfreiheit zurückgeführt wird [27].

Die mammalischen Netzwerke, denen es an einer skalenfreien Architektur mangelt, verfügen dennoch auch über eine Störanfälligkeit, die sich in einer generellen Robustheit mit einer teilweise hohen Sensitivität äußert. Nur wenige Knoten weisen einen PDI auf, der den mit 0.012 (Transkription), 0.0023 (Signalübertragung) und 0.0019 (Metabolismus) kleinen durchschnittlichen PDI deutlich übersteigt (Abbildung 5). Den Hubknoten lässt sich jedoch nicht die Schlüsselposition für die Knoten-zu-Knoten Verbindungen zuschreiben: Sie sind nur zum Teil von großer topologischer Signifikanz (z.B. das Gen des Enzyms *Nucleoside-diphosphate kinase* im metabolischen Netzwerk), viele Hubknoten, wie etwa der G-Protein Komplex *Gi* im Signalübertragungsnetzwerk, haben sogar einen sehr niedrigen PDI. Dagegen spielen vielmehr auch Knoten mit kleinem Degree eine große Rolle für die Knoten-zu-Knoten Verbindungen, u.a. mit *IRF-1*, welches in der Immunantwort und Apoptose eine wichtige Funktion ausübt (Transkription). Die Sensitivität der mammalischen Netzwerke hängt damit nicht mit der Stärke der lokalen Vernetztheit eines Knotens zusammen. Für andere biologische Netzwerke, bei denen sich die Annahme eines skalenfreien

Aufbaus ebenfalls nicht bestätigt, ist dies noch zu überprüfen.

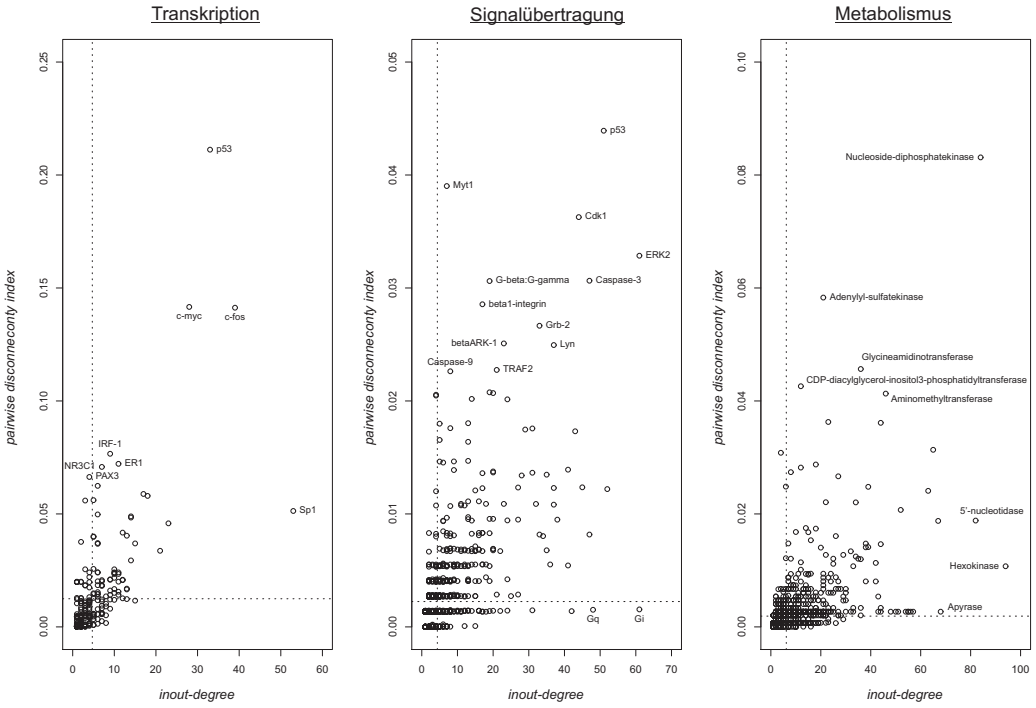


Abbildung 5: *Pairwise disconnectivity index* versus *inout-degree* in den mammalischen Netzwerken. Die jeweiligen Mittelwerte sind durch die gestrichelten Linien gekennzeichnet.

Darüber hinaus wird aus Abbildung 5 auch deutlich, dass sich die mammalischen Netzwerke in ihrer Störanfälligkeit unterscheiden: Im Transkriptionsnetzwerk hängt im Mittel die Verbindung zwischen fünfmal mehr Knotenpaaren von der Präsenz eines Knotens ab, als in den anderen beiden Netzwerken[16]. Des weiteren variiert sowohl die Anzahl an topologisch signifikanten Knoten, wie auch die Stärke ihres Einflusses, erheblich zwischen den Netzwerken. Vor allem haben im Transkriptionsnetzwerk nominell wenige Gene vergleichsweise sehr hohe PDI-Werte[17]. Das lässt auf einen zentrierteren Aufbau des Transkriptionsnetzwerks schließen, bei dem einzelne transkriptionsfaktorkodierende Gene einen Großteil der Verbindungswege kontrollieren und das System anfälliger gegenüber gezielten Knockouts machen. Genau umgekehrt verhält es sich in den anderen beiden Netzwerken: Die Gliederung in verschiedene, ineinander übergreifende Pathways

[16] Auf Grundlage der durchschnittlichen PDI-Werte in den Netzwerken (siehe vorherigen Abschnitt).

[17] Beispielsweise hat das Gen des Tumorsupressors *p53* im Transkriptionsnetzwerk mit einem PDI von $0,21$ den größten PDI, der im Signalnetzwerk nur bei $0,04$ (*p53*) und im metabolischen Netzwerk bei $0,08$ (*Nucleoside-diphosphate kinase*) liegt.

(Signalkaskaden bzw. Enzymwege) bietet wesentlich mehr Alternativpfade, wovon Knoten-zu-Knoten Verbindungen grundsätzlich profitieren. Engpässe in Form von Signalproteinen bzw. enzymkodierenden Genen üben deswegen einen geringeren Einfluss aus.

6 Motive und andere topologische Muster

Der modulare Aufbau biologischer Systeme wird gegenwärtig als effektiver Mechanismus für eine robuste, aber dennoch sehr anpassungsfähige Architektur angesehen [56]. Module sind in erster Linie Ansammlungen von Genen, Proteinen, etc., die gemeinschaftlich eine bestimmte Funktion ausüben, nach der sie sich prinzipiell voneinander abgrenzen lassen [57, 58]. Eine zentrale Rolle wird im Rahmen von Modulen den sogenannten Motiven (englisch: *motifs*) beigemessen, die als die einfachsten Bausteine angesehen werden, aus denen die größeren Module in biologischen Systemen bestehen [1, 50, 57].

Ein Motiv ist ein in einem Netzwerk außergewöhnlich oft vorkommendes topologisches Muster (englisch: *pattern*), verglichen mit seinem Auftreten in zufälligen Netzwerken von gleicher Größe und *degree*-Verteilung. Topologische Muster[18] beschreiben den jeweiligen Zusammenhang einer definierten Menge miteinander verknüpfter Knoten, der sich aus der Gesamtheit aller Kanten zwischen diesen Knoten ergibt. Im Vordergrund steht daher mehr die gemeinsame Struktur von einigen Subgraphen als die konkreten Knoten und Kanten. Motive werden üblicherweise anhand des Z-Score Wertes ermittelt, der angibt, ob die Häufigkeit des Auftretens eines Musters in einem realen Netzwerk sein mittleres Vorkommen in einer großen Anzahl randomisierter Netzwerke übersteigt [59]. Kann dies für ein Muster beobachtet werden, so liegt ein positiver Z-Score vor und das Muster wird als Motiv gekennzeichnet.

Besondere Aufmerksamkeit erfahren unter den Mustern die Motive, weil mit ihrer Überrepräsentation eine grundlegende, funktionale Bedeutung und Präferenz, das jeweilige Motiv während der Evolution des betrachteten Systems zu erhalten, verbunden wird. Von Proteinen, die Bestandteil mehrerer Motive im Proteininteraktionsnetzwerk von *S. cerevisiae* sind, wird beispielsweise vermutet, in weiteren Spezies hochkonserviert zu sein [60, 61]. Die gleichen Motive wurden zudem in anderen Organismen von Bakterien über Pflanzen bis hin zu Tieren gefunden [62]. In biologischen Systemen taucht vor allem der 3-Knoten *feed-forward loop* als Motiv auf (Tabelle 1, Muster mit ID 38), dem eine Reihe regulatorischer Vorgänge entsprechen [63, 64, 65]. Dabei handelt es sich z.B. um Signal-sensitive Prozesse, die aufgrund eines Stimulus entweder eine sofortige oder verzögerte Reaktion auslösen.

[18] Im folgenden wird nur der Ausdruck *Muster* verwendet.

		Transkription		Signalübertragung		Metabolismus	
Muster	ID	Hfk.	Z-Score	Hfk.	Z-Score	Hfk.	Z-Score
	6	1916	-0.39	11774	-8.92	10390	-82.38
	12	1068	-1.67	11865	-9.67	14208	-44.75
	14	73	-10.47	881	-9.48	1118	-34.37
	36	1620	-2.91	13606	-6.91	47485	102.03
	38	129	5.68	496	14.24	1627	68.15
	46	17	11.18	29	8.76	85	27.47
	78	4	-10.85	49	-5.41	336	-69.02
	102	3	0.35	23	8.78	101	29.04
	140	1	-0.88	38	4.64	29	6.27
	164	197	-6.52	722	-10.15	4228	-69.65
	166	20	7.12	21	8.38	492	82.09
	174	6	7.31	9	6.83	71	23.75
	238	1	-	-	-	49	729.30

Tabelle 1: Häufigkeiten und Z-Score Werte der dreizehn möglichen Muster aus drei Knoten in den mammalischen Netzwerken. Die Spalte *Muster* skizziert die Struktur eines Musters, die zugehörige Identifkationsnummer ist in Spalte *ID* angegeben. Die Spalte *Hfk.* führt an, wie oft ein Muster zwischen je drei verschiedenen Knoten in einem Netzwerk vorkommt. Ein positiver *Z-Score* zeigt die Überrepräsentation eines Musters an (rot markierte Einträge).

Neben dem *feed-forward loop* sind Motive in den mammalischen Netzwerken vorzugsweise unter denjenigen Mustern aus drei Knoten zu finden, die einem vorwärts- und/oder rückwärtsgerichteten (*feed-forward*, *feedback*) *loop* bilden (Tabelle 1). *Feedback loops* können als „*Switch*“ auftreten, um die Aktivitätszustände von Proteinen zu verändern, wie z.B. das Blocken der extrazellulären Bindestelle eines Rezeptors nach dem Einschleusen eines Signals. Interessanterweise ist der vor allem für Signalkreisläufe und Stoffwechselprozesse typische *feedback loop* (ID 140 in Tabelle 1) [66, 67, 68, 69, 70] lediglich in geringer Anzahl im mammalischen Transkriptionsnetzwerk vorhanden[19]. Muster, die sich nur aus wenigen Kanten zusammensetzen und bei denen nicht alle Knoten miteinander

[19] Dasselbe gilt für den *feedback* Mechanismus mit ID 73.

verknüpft sind, kommen mit weitem Abstand am häufigsten vor, sind aber i.d.R. mit einem negativem Z-Score versehen.

Die Fokussierung auf Motive und die ihnen zugesprochene funktionale Bedeutung ist allerdings umstritten. Zum einen ist ihre einwandfreie Detektion problematisch, weil sich die Überrepräsentation eines Musters nicht zuverlässig reproduzieren lässt [18, 71]. Zum anderen ist es fraglich, ob überhaupt aus dem ungewöhnlichen Vorkommen eines Musters auf eine besondere funktionale und/oder strukturelle Bedeutung geschlossen werden kann. Motive müssen nicht notwendiger Weise die evolutionär selektierten Bausteine biologischer Systeme sein und eine generell wichtige Rolle bei regulatorischen Vorgängen einnehmen [72, 73]. Möglicherweise sind solche Eigenschaften aber bei Mustern unabhängig von ihrer Häufigkeit zu verzeichnen oder sogar nur spezifischen Kombinationen von Genen/Proteinen zuzuschreiben. Zwar wird eine enge topologische Beziehung zwischen den globalen Charakteristika komplexer Netzwerke und den enthaltenen Mustern erwartet, jedoch ist sie bislang wenig untersucht und verstanden worden [74]. Der Grund hierfür liegt darin, dass sich die meisten Analysen auf die kinetischen Eigenschaften von Motiven konzentrieren und auf die Art und Weise, wie sie informationsverarbeitende Funktionen ausüben [62, 64, 65, 75]. Ihre Aussagekraft ist, bedingt durch die Fokussierung auf die interne Struktur von Motiven, jedoch lediglich auf eine lokale Ebene beschränkt.

Um die globale Bedeutung von Motiven für die Topologie eines Netzwerks abschätzen zu können, sollten sie in seinem Gesamtkontext zusammen mit den anderen, nicht überrepräsentierten Mustern betrachtet werden. Dabei gilt es zu bedenken, dass Muster unabhängig ihres Vorkommens in erster Linie Klassifizierungsmerkmale für die prinzipiell möglichen Verknüpfungsstrukturen in gerichteten Graphen sind, die sich aus den Kanten zwischen einer Anzahl zusammenhängender Knoten ergeben können. Demzufolge unterscheidet sich das i-te Muster aus n Knoten, P_i^n, von den weiteren Mustern mit gleicher Knotenanzahl einerseits darin, zwischen welchen der n Knoten Kanten bestehen und andererseits in der Menge seiner Instanzen im gerichteten Graphen G (Abbildung 6). Die j-te Instanz des i-ten Musters, $P_{i,j}^n$, ist ein Subgraph von G, der aus der einmaligen Kombination von n Knoten sowie allen Kanten, die zwischen diesen Knoten in G existieren, besteht, und dessen Verknüpfungsstruktur mit dem des Musters P_i^n identisch ist[20]. Zwangsläufig ergibt sich damit die globale Bedeutung

[20] Es gibt daher keine weitere Instanz des i-ten oder eines anderen Musters, die sich aus genau den gleichen n Knoten zusammensetzt wie $P_{i,j}^n$.

eines Musters für ein Netzwerk aus der seiner einzelnen Instanzen.

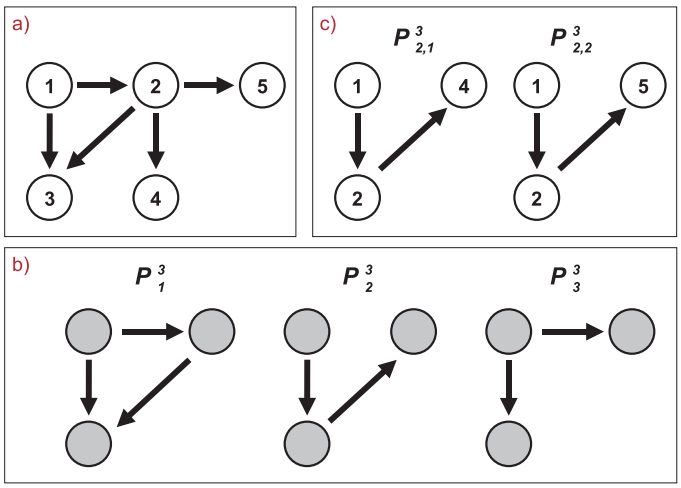


Abbildung 6: Muster und ihre Instanzen. **a)** Ein gerichteter Graph G mit den Knoten $V = \{1, 2, \cdots, 5\}$. **b)** Die Menge aller verschiedenen Muster aus 3 Knoten in G. **c)** Alle Instanzen des Musters P^3_2.

Mit dem *pairwise disconnectivity index* ist in Kapitel 4 ein Ansatz vorgestellt worden, der sich auch zur Bemessung der topologischen Signifikanz einer Muster-Instanz verwenden lässt [46]. Ausgehend von der gleichen Grundidee geht es hier darum zu quantifizieren, wie essentiell die Verknüpfungsstruktur zwischen den n Knoten einer Muster-Instanz für die Konnektivität eines Netzwerks ist. Dies lässt sich erreichen, indem alle Kanten zwischen den Knoten von $P^n_{i,j}$ eliminiert und der sich daraus ergebende Effekt auf die Anzahl der verbundenen, geordneten Knotenpaare in G gemessen wird. Je mehr der Knoten-zu-Knoten Verbindungen dadurch vollständig unterbrochen wurden, umso größer ist die topologische Signifikanz von $P^n_{i,j}$. Der *pairwise disconnectivity index* der Muster-Instanz $P^n_{i,j}$ ist definiert als

$$PDI(P^n_{i,j}) = 1 - \frac{N'}{N} \tag{6.1}$$

und gibt an, wie sich die Anzahl N an verbundenen, geordneten Knotenpaaren im Graphen G von der Anzahl N' im Graphen G' verändert. Dabei unterscheidet sich der Graph G' von G nur durch das Fehlen der Kanten, die zwischen den Knoten von $P^n_{i,j}$ in G existieren; die Knoten selber bleiben erhalten. Für die Bewertung der topologischen Signifikanz des zugehörigen Musters P^n_i sind neben dem PDI der Instanz $P^n_{i,j}$ entsprechend alle weiteren PDI-Werte seiner insgesamt J Instanzen

gleichermaßen relevant, so dass mit

$$PDI(P_i^n) = \frac{1}{J}\sum_{j=1}^{J} PDI(P_{i,j}^n) \tag{6.2}$$

die Bedeutung des Musters P_i^n für die Konnektivität von G als der mittlere Einfluß auf die Knoten-zu-Knoten Verbindungen eines Subgraphen aus n Knoten, dessen Verknüpfungsstruktur dem des Musters P_i^n entspricht, verstanden werden kann.

Für die mammalischen Netzwerke zeigte sich, dass dieser Einfluß für Muster aus drei Knoten generell als sehr gering einzuschätzen ist und es auch keinen Unterschied macht, ob ein Muster überrepräsentiert ist oder nicht [76]. Im Transkriptionsnetzwerk werden durchschnittlich weniger als 1% der Knoten-zu-Knoten Verbindungen durch das Löschen der Instanzen eines Musters unterbrochen; in den anderen beiden Netzwerken sind die PDI-Werte sogar noch um eine Größenordnung kleiner (Abbildung 7). Wesentlich empfindlicher, als

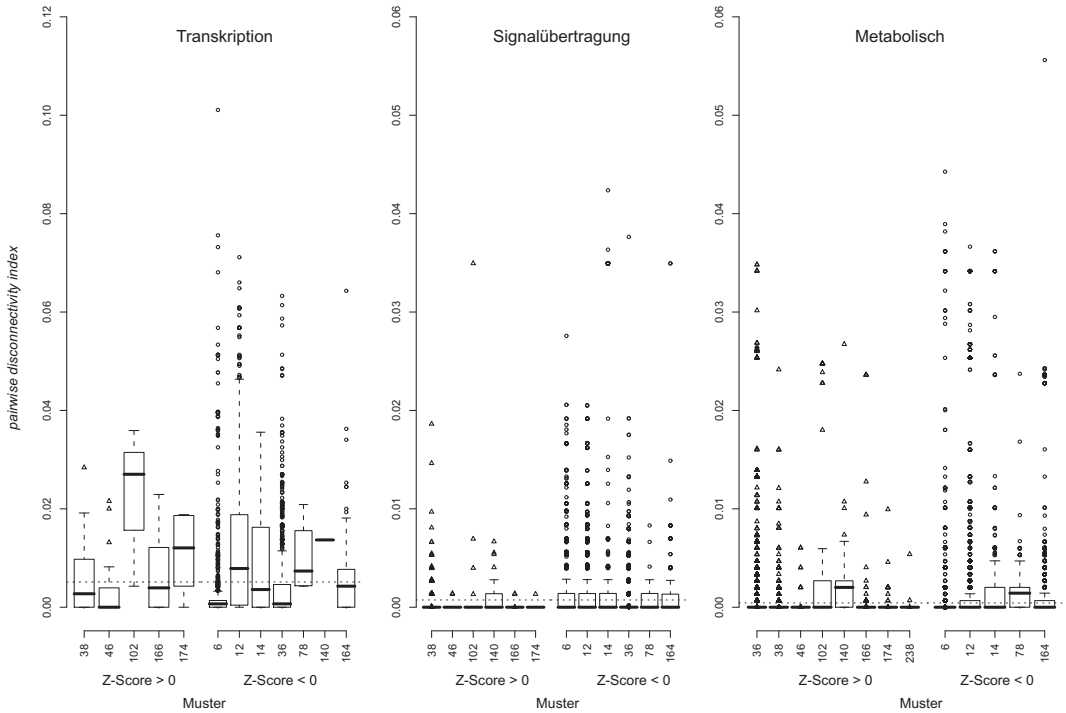


Abbildung 7: Die topologische Signifikanz von Muster-Instanzen aus drei Knoten. Die Boxplots zeigen den PDI der Muster-Instanzen, aufgeteilt jeweils nach den Mustern mit positiven (linke Seite) und negativem (rechte Seite) Z-Score.

diese kleinen PDI-Werte zunächst vermuten lassen, reagieren die Netzwerke nur

gegenüber dem gezielten Entfernen einzelner Instanzen. Beispielsweise haben im metabolischen Netzwerk 75% der Muster-Instanzen einen PDI von 0, während gerade 0,7% von ihnen zwischen einem und acht Prozent der Knoten-zu-Knoten Verbindungen vermitteln[21]. Eine präferierte Zugehörigkeit zu den Motiven fehlt jedoch auch bei der geringen Menge an Instanzen mit größeren PDI-Werten: Können im metabolischen Netzwerk solche Instanzen noch nahezu allen Mustern zugeordnet werden, gehören sie im Transkriptions- und Signalübertragungsnetzwerk eher zu den Mustern mit negativen Z-Score.

Für die Konnektivität der mammalischen Netzwerke spielen demnach nur spezifische Kombinationen aus jeweils drei Knoten eine wichtige Rolle. Können diese Instanzen auch keinem bestimmten Muster zugeordnet werden, scheint es dafür eine enge funktionale Beziehung zu geben. Zum Beispiel setzen sich die Muster-Instanzen mit den jeweils zehn größten PDI-Werten in den mammalischen Netzwerken durchweg aus Genen/Proteinen zusammen, die an elementaren Prozessen beteiligt sind und zum Teil überlebenswichtige Aufgaben erfüllen. Handelt es sich im Transkriptionsnetzwerk um Gene, die vor allem Zellproliferation und -differenzierung steuern (z.B. *Mitf*), sind es im Signalnetzwerk bekannte Zellzyklusregulatoren (u.a. *p53*, *Cdk1*) und im metabolischen Netzwerk Enzyme, die als Katalysatoren bei der Energiegewinnung oder Synthese von Zellwachstumsstoffen/Sekundären Botenstoffen agieren. Des Weiteren haben diese Muster-Instanzen eine auffallend große Übereinstimmung in den Knoten und Kanten aus denen sie bestehen, so dass sie statt voneinander getrennter Subgraphen miteinander verbundene Komponenten bilden[22]. Für das Transkriptions- und Signalübertragungsnetzwerk hat sich herausgestellt, dass die Muster-Instanzen zu einer einzigen Zusammenhangskomponente miteinander verknüpft sind, die im Fall des Signalübertragungsnetzwerks große Teile der ablaufenden Vorgänge bei der Initiierung der Apoptose während der Mitose [77] beschreibt.

Auch für andere Spezies ließ die Analyse von Mustern aus drei Knoten eine funktionale Beziehung bei den wenigen Instanzen erkennen, für die, entgegen

[21] Im Transkriptions- und Signalübertragungsnetzwerk haben 30% bzw. 50% einen PDI von 0 und ca. 15% bzw. 0,5% können jeweils 1% der Knoten-zu-Knoten Verbindungen unterbrechen.

[22] Unter anderem teilen sich fünf der zehn Muster-Instanzen aus dem Transkriptionsnetzwerk die Kante *c-myc* → *PAX3*, im Signalnetzwerk sogar acht die Kanten *Cdk1* ↔ *Myt1* und im metabolischen Netzwerk sieben die Kante *Nucleoside-diphosphatekinase* → *UDP-GlcNAc pyrophosphorylase*. Die Vermutung, dass die Bedeutung der Muster-Instanzen für die Knoten-zu-Knoten Verbindungen daher auf einzelne Kanten zurückzuführen ist, hat sich allerdings nicht bewahrheitet [46].

dem allgemeinen Trend, ein größerer Einfluß auf die Konnektivität gemessen wurde [46]. Keine Unterstützung fand sich wiederum bei den untersuchten Transkriptionsnetzwerken von *E. coli* und *S. cerevisiae* für eine mögliche Sonderrolle von Motiven in diesem Zusammenhang, obwohl sich die Netzwerke wegen der einfacheren Transkriptionsmechanismen in den beiden Spezies durch deutlich weniger komplexe Muster als das mammalische Transkriptionsnetzwerk auszeichnen[23].

[23] In *S. cerevisiae* und *E. coli* kommen nur sieben bzw. vier der dreizehn möglichen Muster überhaupt vor, von denen vier bzw. zwei überrepräsentiert sind. Bis auf den *feed-forward loop*, der das einzig gemeinsame Motiv mit den mammalischen Netzwerken ist, sind *loops* quasi Fehlanzeige.

7 Autoregulation

In den mammalischen Netzwerken bilden die autoregulatorischen Knoten eine Minderheit[24], deren gehäuftes Vorkommen in den Mustern aus drei Knoten jedoch darauf hindeutet, dass sie möglicherweise ein wesentlicher Bestandteil von Mustern sind. Autoregulatorische Knoten, die im Vergleich zu anderen Knoten außerdem einen durchschnittlich höheren *degree* und *pairwise disconnectivity index* haben, wurden im Transkriptionsnetzwerk bei über 95% und in den anderen beiden Netzwerken bei mehr als der Hälfte der Muster-Instanzen aus drei Knoten gefunden [76]. Weitere Analysen haben ergeben, dass im Transkriptionsnetzwerk sogar ein sehr hoher Anteil der Muster-Instanzen auch zwei autoregulatorische Knoten hat. In diesem Zusammenhang tritt die gegenseitige Regulation zweier autoregulatorischer Gene, der sogenannte Binärloop, besonders häufig auf.

[24] Ihr Anteil an der Gesamtmenge an Knoten beträgt 30% (Transkription), 3,5% (Signalübertragung) und 3,5% (Metabolismus).

8 Fazit und Ausblick

In der vorliegenden Arbeit konnten zum ersten Mal die Prozesse der Transkriptionsregulation, der Signaltransduktion und des Metabolismus in höheren Eukaryoten systemweit untersucht werden. Verwirklicht werden konnte dies durch das Zusammentragen experimentell verifizierter molekularer Interaktionen zu regulatorischen Netzwerken in Form gerichteter Graphen. Als sehr praktikabel hat sich für die Modellierung die Technik der orthologen Abstraktion erwiesen, bei der Daten aus den Spezies Mensch, Maus und Ratte auf Grundlage der funktionellen Ähnlichkeit ihrer Gene verarbeitet und zu einer spezies- und zellübergreifenden Sichtweise gebündelt wurden. Dadurch ließen sich vorhandene Wissensdefizite über regulatorische Vorgänge in den einzelnen Spezies und Zelltypen ausgleichen und aussagekräftige Modelle erzeugen, die sich auf die Kausalzusammenhänge zwischen Genen/Proteinen konzentrieren.

In Bezug auf die strukturellen Eigenschaften der mammalischen Netzwerke konnte ihre topologische Analyse die weitgehend beobachtete Abweichung von zufälligen Netzwerken bestätigen. Diese äußert sich vor allem in der Art und Weise, wie die durch Knoten abstrahierten Transkriptionsfaktoren, Signalproteine und metabolischen Enzyme direkt einander beeinflussen. Im Gegensatz zu Zufallsnetzwerken, bei denen sich die Anzahl an Interaktionspartnern pro Knoten auf einen bestimmten Wert einpendelt, sinkt sie deutlich mit zunehmender Menge möglicher Interaktionspartner. Dieser Trend gilt sowohl für den Umfang, in dem Transkriptionsfaktoren, Signalproteine und metabolische Enzyme reguliert werden als auch für ihren eigenen Wirkungsgrad.

Die *degree*-Verteilungen der mammalischen Netzwerke ließen sich jedoch nicht durch das für Skalenfreiheit erforderliche *power-law* Modell approximieren, sondern zeigten eine gute Übereinstimmung mit Modellen, die einen exponentiellen Verlauf berücksichtigen. Die in den letzten Jahren geäußerte Kritik an der Universalität der Skalenfreiheit hat sich damit als berechtigt erwiesen. Dabei ist Skalenfreiheit ohnehin nur schwer mit der Evolvierbarkeit biologischer Systeme vereinbar, weil die, skalenfreien Netzwerken implizite, homogene Architektur wenig Variabilität beim Wachstum zulässt. Aus diesem Grund ist auch ein alleiniges Wachstum nach dem Mechanismus des *preferential attachment*, der einen *power-law* verteilten *degree* generiert, unwahrscheinlich. Alternativ vorgeschlagene Wachstumsprinzipien, wie Genduplikation und -diversifikation [78, 79], bieten zwar vielversprechende Ansätze, können die Topologie biologischer Netzwerke

aber ebenso wenig vollständig erklären [36]. In dieser Richtung ist deshalb noch ein großer Forschungsbedarf vorhanden.

Robustheit gegenüber Störungen ist dagegen auch für die Architektur der mammalischen Netzwerke charakteristisch und in der Tat ein elementares Aufbauprinzip evolvierender Systeme. Es bedarf jedoch gewisser Abstriche in der Robustheit, um auch Variabilität zu ermöglichen, weshalb die enorme Widerstandskraft der mammalischen Netzwerke gegenüber willkürlichen Störungen einer verhältnismäßig großen Sensitivität weicht, wenn gezielt bestimmte Gene bzw. Proteine ausgeschaltet werden. In dieser Hinsicht hat sich die Regulation transkriptionsfaktorkodierender Gene in höheren Eukaryoten als wesentlich empfindlicherer Kreislauf gegenüber der Steuerung signalverarbeitender oder metabolischer Prozesse erwiesen, was die Frage aufwirft, ob dieser Unterschied auch in andereren Spezies (z.B. Einzeller) zu verzeichnen ist. Entgegen der weitverbreiteten Ansicht, dass die Hubknoten biologische Netzwerke störanfällig machen, deuten die Ergebnisse dieser Arbeit daraufhin, dass ihre Sensitivität nicht mit der Stärke der lokalen Vernetztheit eines Knotens zusammenhängt. Stattdessen hat sich der Beitrag eines Knotens zum Erhalt der Konnektivität eines Netzwerks als zutreffendes Kriterium herausgestellt.

Die Leistungsfähigkeit des *pairwise disconnectivity index*, der diesen Beurteilungsmaßstab nutzt, konnte an diversen Netzwerken aus mehreren Spezies zur Identifikation zentraler Knoten mit experimentell nachgewiesenen Schlüsselfunktionalitäten der jeweiligen Genprodukte demonstriert werden. Damit steht ein neues Verfahren zur Analyse regulatorischer Systeme zur Verfügung, dessen Arbeitsweise leicht nachvollziehbar ist und das unabhängig von der Semantik des zu untersuchenden Netzwerks eingesetzt werden kann. Die Methodik zeichnet sich zudem durch die einfache Interpretierbarkeit der Ergebnisse und ihre Vielseitigkeit aus, wie durch die Übertragung des Ansatzes auf topologische Muster veranschaulicht werden konnte. Ein Java Programm [48] und ein Webservice [47, 49] wurden im Rahmnen dieser Arbeit implementiert und stehen für die systematische Analyse weiterer biologischer Netzwerke mit dem *pairwise disconnectivity index* zur Verfügung.

Das vielbeachtete Konzept der Motive, welches statistisch überrepräsentierte topologische Muster hervorhebt, hat zu Recht darauf aufmerksam gemacht, wie Regulation typischerweise im kleinen Maßstab funktioniert. Dass sie in höheren Eukaryoten komplexer als in einfacheren Organismen ist, konnte aus der

größeren Vielfalt an topologischen Mustern - speziell von denjenigen mit einer *loop* Struktur - unabhängig von ihrem Vorkommen abgelesen werden. Es besteht zudem Anlass zur Vermutung, dass autoregulatorische Gene/Proteine wichtige Einflußfaktoren der hinter den Mustern stehenden regulatorischen Vorgänge sind. An der Ausnahmestellung von Motiven als evolutionär selektierte Interaktionsmuster mit besonderer funktionaler Relevanz für das jeweilige Gesamtnetzwerk darf allerdings stark gezweifelt werden. Es wurden keine Hinweise gefunden, dass Motive im globalen Kontext regulatorischer Netzwerke eine tragende Rolle spielen oder von größerem Stellenwert sind als andere topologische Muster. Eine Beziehung zwischen der Konnektivität der Netzwerke und Mustern, die aus drei Knoten bestehen, hat sich nur für wenige Muster-Instanzen herausgestellt, deren Bedeutung sich auf den spezifischen funktionalen Zusammenhang der jeweiligen Kombination aus Genen/Proteinen zurückführen ließ.

In der Analyse solcher regulatorischen Module, insbesondere derer größeren Umfangs, steckt noch ein großes Potenzial, da sie darüber Aufschluss geben kann, wie bestimmte zelluläre Abläufe durch den Zusammenschluss einzelner Gene/Proteine ausgeführt werden. Interesse besteht hierbei nicht nur in der Erkennung der Module, sondern auch darin, wie die Schnittstellen zwischen ihnen beschaffen sind und sich die großen Module in kleinere zerlegen lassen. Eine weitere Herausforderung ist die Vervollständigung der in dieser Arbeit untersuchten regulatorischen Systeme. Hierfür kann ergänzend zur Durchführung von Laborexperimenten die Möglichkeit in Erwägung gezogen werden, Methoden zur Vorhersage molekularer Interaktionen zu verwenden. Diese sind im Vergleich weitaus weniger aufwendig und restriktiv hinsichtlich ihres Einsatzes; ihre Vorhersagegüte ist zum gegenwärtigen Zeitpunkt allerdings noch ausbaufähig. Als Benchmark zur Überprüfung der Qualität von Netzwerken, die prognostizierte Informationen beinhalten, können aber u.a. die in dieser Arbeit ermittelten topologischen Eigenschaften genutzt werden.

Anhang A

Abbildungen

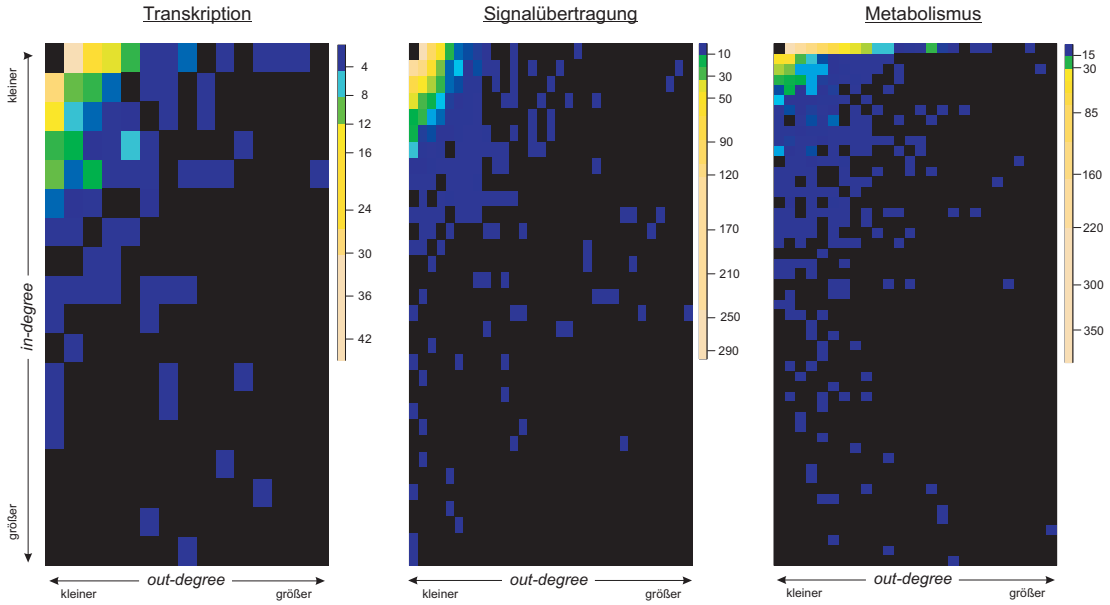


Abbildung A.1: Häufigkeitsverteilung der möglichen Kombinationen von *in-* und *out-degree* eines Knotens. Jede *heatmap* zeigt, wie viele der Knoten in den mammalischen Netzwerken einen bestimmten *in-* und *out-degree* haben. Während die meisten Knoten einen kleinen *in-* und *out-degree* haben (beige Farbbereiche), kommen viele der möglichen Kombinationen mit großem *in-* oder *out-degree* nur sehr wenig (lila Bereiche) oder gar nicht vor (schwarze Bereiche). Vor allem sind Hubknoten damit eher durch einen großen *in-* und kleinen *out-degree* gekennzeichnet et vice versa.

Anhang B

Artikel 1 aus dem Jahr 2008 in BMC Bioinformatics

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks

Anatolij P Potapov*[1], Björn Goemann[1] and Edgar Wingender[1,2]

Address: [1]Department of Bioinformatics, Medical School, Georg August University of Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany and [2]BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbüttel, Germany

Email: Anatolij P Potapov* - anatolij.potapov@bioinf.med.uni-goettingen.de; Björn Goemann - bjoern.goemann@bioinf.med.uni-goettingen.de; Edgar Wingender - e.wingender@med.uni-goettingen.de

* Corresponding author

Published: 2 May 2008

BMC Bioinformatics 2008, **9**:227 doi:10.1186/1471-2105-9-227

Received: 23 July 2007
Accepted: 2 May 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/227

Abstract

Background: Currently, there is a gap between purely theoretical studies of the topology of large bioregulatory networks and the practical traditions and interests of experimentalists. While the theoretical approaches emphasize the global characterization of regulatory systems, the practical approaches focus on the role of distinct molecules and genes in regulation. To bridge the gap between these opposite approaches, one needs to combine 'general' with 'particular' properties and translate abstract topological features of large systems into testable functional characteristics of individual components. Here, we propose a new topological parameter – the pairwise disconnectivity index of a network's element – that is capable of such bridging.

Results: The pairwise disconnectivity index quantifies how crucial an individual element is for sustaining the communication ability between connected pairs of vertices in a network that is displayed as a directed graph. Such an element might be a vertex (i.e., molecules, genes), an edge (i.e., reactions, interactions), as well as a group of vertices and/or edges. The index can be viewed as a measure of topological redundancy of regulatory paths which connect different parts of a given network and as a measure of sensitivity (robustness) of this network to the presence (absence) of each individual element. Accordingly, we introduce the notion of a path-degree of a vertex in terms of its corresponding incoming, outgoing and mediated paths, respectively. The pairwise disconnectivity index has been applied to the analysis of several regulatory networks from various organisms. The importance of an individual vertex or edge for the coherence of the network is determined by the particular position of the given element in the whole network.

Conclusion: Our approach enables to evaluate the effect of removing each element (i.e., vertex, edge, or their combinations) from a network. The greatest potential value of this approach is its ability to systematically analyze the role of every element, as well as groups of elements, in a regulatory network.

Background

Recent advances in graph theory have provided a new view on the topological design of different real-world networks [1-6]. Such systems exhibit small-world properties: They are surprisingly compact (i.e., their diameter is rather small) and display increased clustering features [7]. Moreover, they show a scale-free topology and follow a power-law type of the degree distribution: most components exhibit only one or two connections, but a few are involved in dozens and function as hubs, thereby providing networks with high robustness against random failures [1-3]. Various biological networks, such as metabolic or protein-protein interaction networks, show a scale-free topology [1,2,5] that emerges as a hallmark of modern systems biology.

However, by itself, the fact that a network has scale-free features is of limited practical use to biologists because power laws occur widely in nature and can have many different origins [8]. Currently, there is a gap between purely theoretical studies of the topology of large regulatory networks, on the one hand, and the practical traditions and interests of experimentalists, on the other hand. While the theoretical approaches emphasize the global characterization of regulatory systems as whole entities, experimental (even high-throughput) approaches usually focus on the role of distinct molecules and genes in regulation. There is a rather limited interface between them. Both approaches have not been integrated to study complex regulatory systems. To reconcile these apparently opposite views, one needs to combine 'general' with 'particular' aspects, as it is attempted by modern systems biology approaches, and translate rather abstract topological features of large systems into testable functional characteristics of individual components. So far, few such graph-theoretical characteristics have been explored for the analysis of biological networks [9-11], which are expected to have their particular properties.

There is a great need for approaches capable to quantitatively evaluate the importance of individual components in complex biological systems. Centrality analysis provides a valuable method for the structural, i.e. topological, analysis of biological networks. It allows to identify key elements within networks and to rank network elements such that experiments can be tailored to interesting candidates [10,11]. Local approaches such as the degree of a vertex (i.e., the number of its adjacent edges) help to find important molecules/genes which directly control many other molecules/genes, but fail to identify key regulators which are capable of affecting other molecules/genes in an indirect fashion. Other parameters, such as closeness and betweenness centrality, consider both local and distant connections within a network [9-12]. Closeness centrality evaluates how close a vertex (molecule/gene) is to all other vertices. Betweenness centrality measures how frequently a vertex appears on all shortest paths between two other vertices in a whole network [12-14]. Liu and colleagues [15] tested relationships between the phylogenetic profile of an enzyme and its topological importance in metabolic networks. They found that betweenness centrality is a good predictor of how many bacterial species have a particular enzyme. In contrast, the relationship with closeness centrality is much weaker or non-existent. This reflects the fact that the closeness centralities of a vertex and its immediate neighbors are rather similar and differ much less than their betweenness centralities. The representative power of betweenness centrality as a biologically relevant parameter was further confirmed in the topological analysis of mammalian networks of transcription factor genes: Among several topological characteristics tested, the betweenness centrality of individual transcription factor genes was found to be the most representative and relevant in regard to the biological significance of distinct elements [16]. In protein networks, betweenness centrality is rather helpful for identifying key connector proteins, i.e., bottlenecks, with particular functional and dynamic properties [17]. Betweenness centrality has been used to search for community structures in biological networks [12] and their hierarchical decomposition into subnetworks [18]. Thus, betweenness centrality has emerged as a promising measure of the biological significance of network elements.

Unfortunately, the approach based on the betweenness centrality suffers from some significant limitations due to the inherent nature of this parameter, which are finally becoming manifested in a restricted qualification for the analysis of regulatory networks. In the following we identify these limitations and propose with the pairwise disconnectivity index a new methodology that overcomes them. Subsequently, we apply the method to the analysis of various biological networks.

Results

Betweenness centrality and its limitations in analyzing regulatory networks

In regard to the needs of an analysis of regulatory networks there are two major disadvantages of betweenness centrality. Firstly, shortest paths are supposed to be the most important ones, which is a big oversimplification and misleading. The importance of a path is determined not so much by its length, i.e., the number of reactions, but rather by the integral efficiency of all these reactions. This efficiency depends on many instances, such as the concentrations of the participants, rate constants, etc. Longer paths can be faster and more efficient than shorter ones. For instance, in regulatory networks, the initiation of transcription and translation is typically governed by sets of specific factors. This increases the length of the cor-

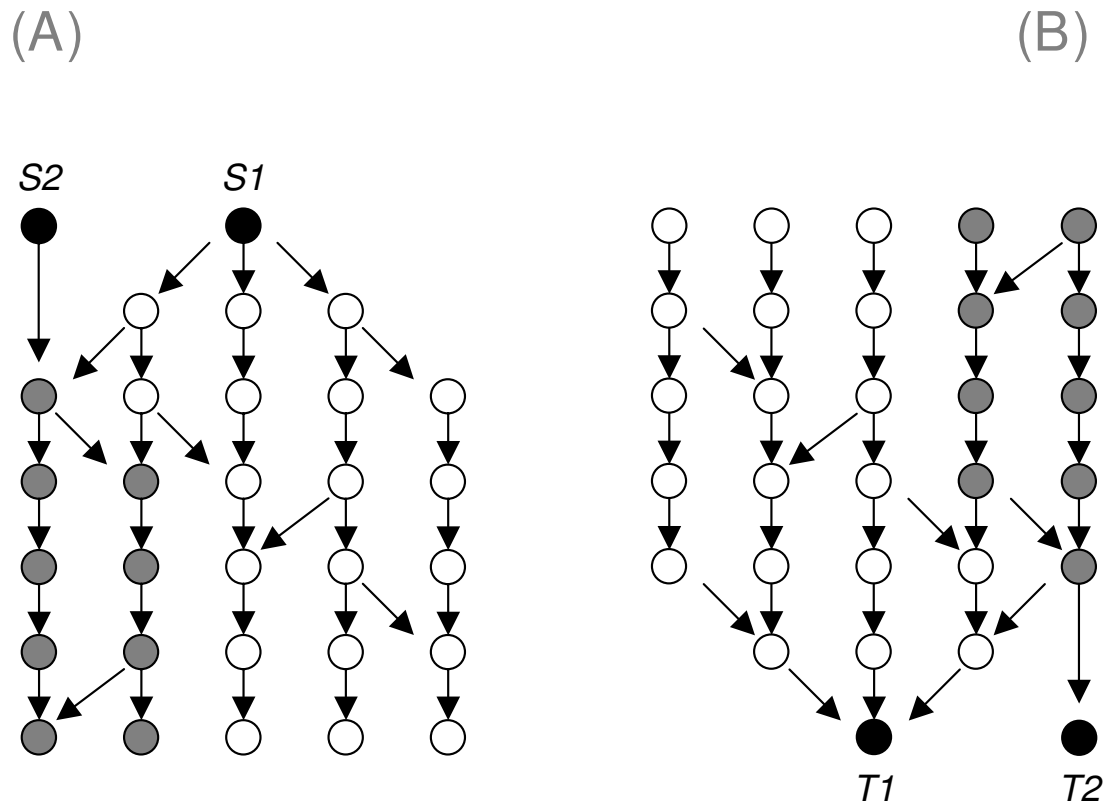


Figure 1
Some vertices at the periphery of a regulatory network (the places where signals start or get their targets) can be rather significant.*A*: The topological impact of start-point *S1* is bigger than that of start-point *S2*. Both, white and gray vertices are on some path beginning in *S1*, while *S2* is limited on the gray ones. ***B***: The topological significance of end-point *T1* is bigger than that of end-point *T2* because of being reachable from all gray and white vertices. However, in terms of betweenness centrality, all of them are attributed with zero values which fail to reflect the individual connectedness of such input/output elements within the whole network.

responding paths, but drastically improves the efficiency and specificity of these processes. In a similar way, scaffold and adaptor proteins, which themselves are not enzymes, recruit downstream effectors in signaling pathways and enhance both the efficiency and specificity of signal propagation. Moreover, in most regulatory networks, like gene networks, an inherent problem is that the real length of edges is not defined at all. Each single edge commonly summarizes a set of events and describes the causal relations between genes. But this kind of abstraction does not say anything about the complexity and length of the corresponding processes. Thus, dealing with inconsistent semantics of the edges renders the definition of a shortest path in these networks highly problematic.

Secondly, betweenness centrality can be applied only to vertices that are between other ones. Peripheral vertices, i.e., vertices having either zero incoming or outgoing degree, are not considered. That immediately excludes many extracellular ligands, receptors, target molecules and genes from the analysis of a signaling network (Figure 1). Such components, however, directly respond to input-output functionality of the network and therefore are of key significance. Moreover, their individual topological significance in the network may vary in a wide range, as it can be seen when comparing the connectedness of the start-points S1 and S2, or end-points T1 and T2 in Figure 1. However, in terms of betweenness centrality, all of them are attributed with zero values which fail to reflect the individual connectedness of such input/output elements within the whole network.

We therefore developed the concept of the pairwise disconnectivity index as a new topological metric, which evaluates alternative though longer paths as well and can be used to characterize the topological significance of all

individual elements in biological regulatory networks. The approach has some similarity to numerical parameters like *vertex-connectivity* or *edge-connectivity* used in graph theory to measure a graph's connectedness [19]. However, our method does not focus on how the removal of distinct elements breaks a given connected graph into disconnected pieces, like the algorithm of Girvan and Newman [12], though a network's disintegration can be considered as well. Instead, our aim was to find a parameter describing more moderate effects in a still connected network.

Topological significance of individual elements in a regulatory network

In a directed graph $G(V,E)$ representing a regulatory network, the vertices $v \in V$ denote biological entities, e.g., proteins, genes, or small molecules. Causal relationships between these entities are made up of directed edges $e \in E$. We denote the *topological significance* of an individual element (vertex, edge or their combination) as how essential for all connections in the network this element is. To quantify this significance we suggest to measure how the elimination of such an element affects the number of connected ordered pairs of vertices. An ordered pair of vertices $\{i, j\} \mid i \neq j$ and $i, j \in V$, is connected iff there is at least one path from vertex i to vertex j in G. Note, that the ordered pair $\{i, j\}$ is different from $\{j, i\}$ in a directed network. The more ordered pairs become disconnected upon the removal of vertex v, the higher is the topological significance of this vertex. We define the *pairwise disconnectivity index of vertex v*, $Dis(v)$, as the fraction of those initially connected pairs of vertices in a network which become disconnected if vertex v is removed from the network

$$Dis(v) = \frac{N_0 - N_{-v}}{N_0} = 1 - \frac{N_{-v}}{N_0} \qquad (1)$$

Here, N_0 is the total number of ordered pairs of vertices in a network that are connected by at least one directed path of any length. It is supposed that $N_0 > 0$, i.e., there exists at least one edge in the network that links two different vertices. N_{-v} is the number of ordered pairs that are still connected after removing vertex v from the network, via alternative paths through other vertices (see vertex 2 in Figure 2B,C). However, the relation of N_{-v} and N_0 conveyed by $Dis(v)$ immediately uncovers the fraction of connected ordered pairs whose communication essentially depends on vertex v. In the extreme case the removal of vertex v destroys all communication in a network resulting in $Dis(v) = 1$. In contrast, $Dis(v) = 0$ refers to a non-crucial vertex which is obviously not connected to any other vertex in a network.

The example presented in Figure 2 also illustrates the difference between the pairwise disconnectivity index and betweenness centrality. Vertex 2 is characterized with equally high (case A) or low (case B) values of both centralities, whereas they largely differ in case C (high betweenness centrality, but low pairwise disconnectivity index). The toy network in Figure 3 further illustrates that betweenness centrality and pairwise disconnectivity index reflect different properties of a vertex in a network. While the vertices 4 and 7 are mediating most of the shortest paths, thereby exhibiting a very high betweenness centrality value, these vertices show a rather low pairwise disconnectivity index since they provide alternative paths. In contrast, vertex 1 displays modest betweenness centrality but has a high topological significance according to its disconnectivity value (Figure 3). Thus, a vertex with high betweenness is not obligatorily topologically significant according to its disconnectivity value. It is only a clue for the fraction of short communication paths between reachable vertices which are provided due to the existence of a particular vertex.

Furthermore, the difference between the pairwise disconnectivity index and betweenness centrality becomes apparent when taking a closer look into the kind of reachable ordered pairs whose connection depends on vertex v. The complete set of those pairs, N_0 - N_{-v}, may include those which are connected by 1) paths that end at vertex v, 2) paths that start at vertex v, and 3) paths that go through vertex v. Other pairs cannot be affected, since they are connected via paths that do not contain any of the edges around vertex v. Accordingly, the pairwise disconnectivity index of vertex v can be represented as follows

$$Dis(v) = \frac{N_0 - N_{-v}}{N_0} = \frac{\sigma_{sv} + \sigma_{st}(v) + \sigma_{vt}}{N_0} \qquad (2)$$

The term $\sigma_{st}(v)$ in Eq. 2 expresses the number of ordered pairs $\{s,t\} \mid s \neq t \neq v$ and $s, t, v \in V$ that are exclusively linked through vertex v. Both, σ_{sv} and σ_{vt} involve v and represent the path-degree of vertex v in terms of all incoming and outgoing paths, respectively. Altogether, $\sigma_{st}(v)$ is not a trivial combination of σ_{sv} and σ_{vt} as Figure 2 shows: Vertex 2 is indeed crucial for connecting vertex 1 to vertex 3 in graph 2A. But in graphs 2B and 2C the same connection $1 \rightarrow 3$ does not depend on vertex 2 anymore, because of the parallel paths. However, vertex 2 still is essential for all paths that start or end in this vertex. The number of such ordered pairs associated with vertex 2, σ_{s2} and σ_{2t}, does not change in the graphs 2A, 2B and 2C, thereby indicating the absence of a simple relationship between the values of σ_{sv}, σ_{vt} and $\sigma_{st}(v)$.

Often one wants to know how many connected pairs $\{i,j\}$ depend on a particular vertex v while disregarding those kinds of pairs that involve the considered vertex, i.e. where

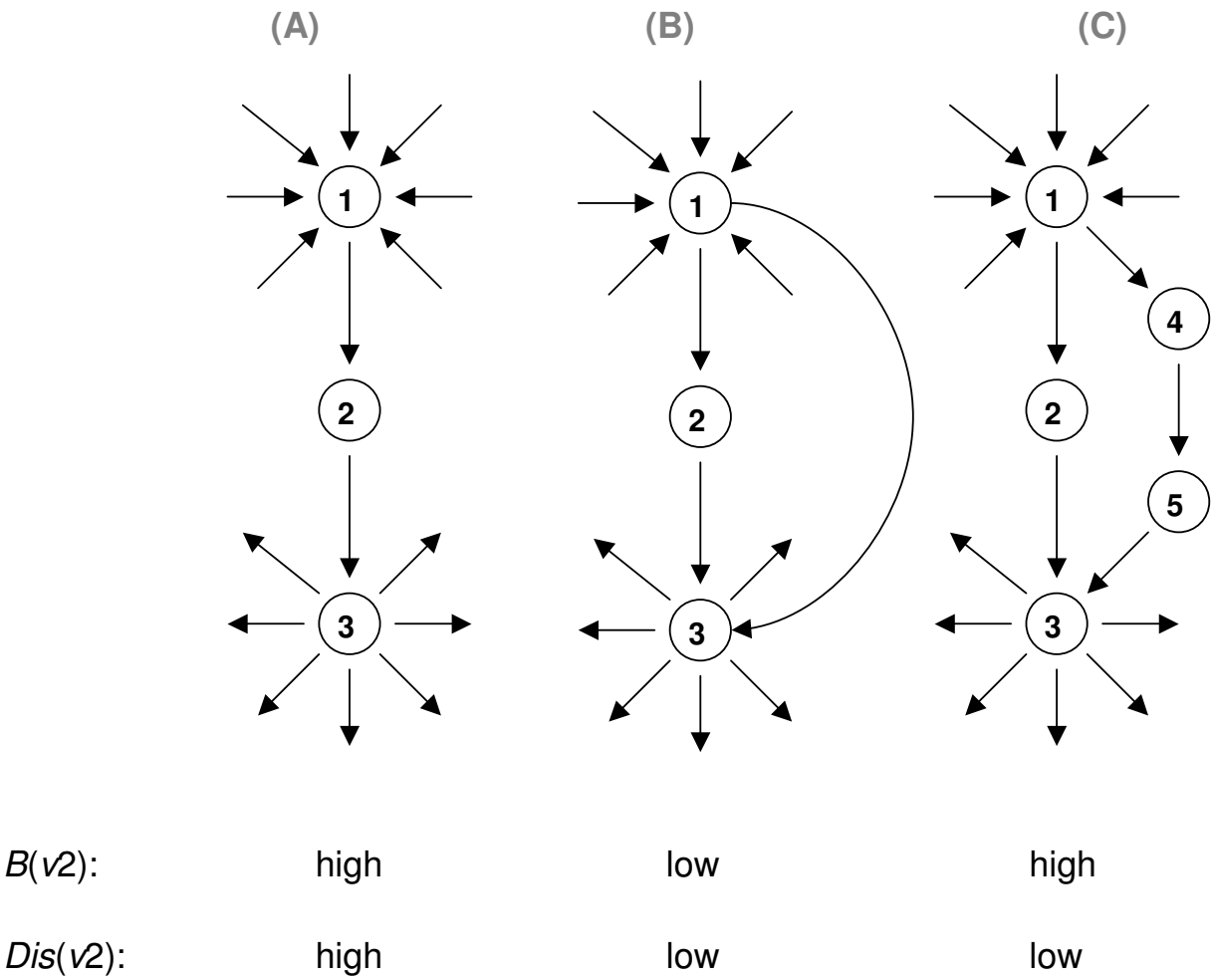


Figure 2
The significance of a vertex is determined by its local and global environments in a network and can be better represented by a set of topological parameters. In cases ***A***, ***B*** and ***C***, the degree of vertex 2 is the same. However, its betweenness centrality in ***A***, ***B*** and ***C*** is high, low and high, respectively, and the parwise disconnectivity index in ***A***, ***B*** and ***C*** is high, low and low, respectively. Thus, focusing on betweenness centrality may yield to misleading conclusions.

$v \neq i$ and $v \neq j$. For example, when analyzing the role of a receptor for the indirect communication of extracellular ligands with transcription factors, communication paths that start or end at the receptor need not to be considered. The term $\sigma_{st}(v)$ in equation 2 exactly comprises this sort of essentiality and we define

$$MDis(v) = \frac{\sigma_{st}(v)}{N_0} \quad (3)$$

as the *mediative disconnectivity index of a vertex v*. It immediately detects the fraction of connected ordered pairs of vertices different from v for whose reachability vertex v is necessary. While the pairwise disconnectivity index of vertex 2 in Figure 2A involves the pairs $\{1,2\}$, $\{1,3\}$ and $\{2,3\}$ it's mediative disconnectivity index reveals that vertex 2 is uniquely bridging the connection for $\{1,3\}$.

The mediative disconnectivity index of a vertex may exhibit some similarity to the beweenness centrality of the vertex. A path that uniquely connects two vertices i and j and is destroyed after removing another vertex v is always the shortest path between i and j. However, betweenness centrality considers all shortest paths and $MDis(v)$ uncovers the cases where vertex v is the only link for a connected pair i and j. The principal difference between these parameters is due to their different sensitivity to the presence of parallel paths: betweenness centrality is insensitive to the

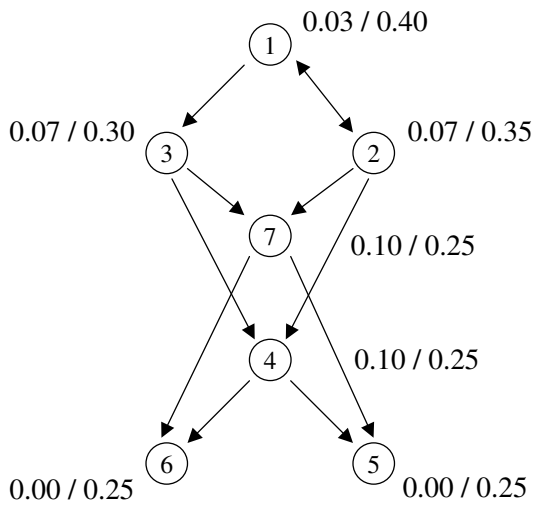


Figure 3
Example for a network which demonstrates that betweenness centrality and pairwise disconnectivity index reflect different properties of a vertex. While the vertices 4 and 7 are mediating many shortest paths, the removal of one of them does not cause a high damage to the existing connections because they provide alternative paths. In contrast, vertex 1 displays a modest betweenness centrality only but has the highest topological significance. The values of the betweenness centrality and the pairwise disconnectivity index of each vertex are indicated as $B(v)$/ $Dis(v)$, respectively.

presence of longer bypasses, whereas $MDis(v)$ is very sensitive to that.

Vertex removal is a strong interference in a network because it simultaneously removes all incoming and outgoing edges of that vertex. One can also perturb a network by selectively knocking out a particular edge. This is a relatively gentle intervention which can simulate various normal and pathological situations in a regulatory network when all components are still present, but due to a mutation in one of them some of its reactions are specifically disabled while others are still working. That is particularly important when considering the fact that edges are a kind of abstraction and simplification, as discussed above. Thus, we declare an edge as topologically significant in the same way as a vertex: The higher the number of ordered pairs that become disconnected the higher the topological significance of an eliminated edge. To quantify this, we introduce the *pairwise disconnectivity index of an edge*, $Dis(e)$, which is defined as

$$Dis(e) = \frac{N_0 - N_{-e}}{N_0} \tag{4}$$

Again, N_0 is the number of ordered pairs of vertices connected by means of at least one directed path in the network. N_{-e} is the number of such pairs after removing edge e from the network. The pairwise disconnectivity index of an edge ranges between $0 \leq Dis(e) \leq 1$. In Figure 2A we previously argued the dependence of the communication of the ordered pair $\{1,3\}$ on vertex 2. With the disconnectivity index of an edge it becomes clear that it is not necessary to remove vertex 2 itself in order to destroy the pair $\{1,3\}$. Moreover, a disorder of either the incoming or outgoing edge of vertex 2 is enough to compass the same effect.

Topological significance of a group of elements in a regulatory network

Not all major functional breakdowns of a network can be explained due to the failure of one single element, but rather to the dysfunction of a subset of vertices or edges. The malfunctioning of this subset may disrupt a significant number of communication lines because parallel paths may be destroyed simultaneously. For example, in Fig. 2C the ordered pair $\{1,3\}$ stays connected unless the vertices 2 and 4 or 2 and 5 are taken out together. As the generalization of Eq. 1 we define the *pairwise disconnectivity index of a group of vertices*, $W \subseteq V$, as

$$Dis(W) = \frac{N_0 - N_{-W}}{N_0} = 1 - \frac{N_{-W}}{N_0} \tag{5}$$

with N_{-W} representing the number of connected ordered pairs after removing the set of vertices W. Note that $Dis(W)$ cannot be inferred directly from the disconnectivity indices of individual vertices in W. This is due to the presence of parallel paths in a network. For example, vertex 4 (or vertex 7) in Figure 3 features a rather low pairwise disconnectivity index. But as part of the group 'vertex 4 AND vertex 7' it causes the network to split into two distinct parts.

Finally, in analogy to Eq. 4 the general case of the removal of an individual edge is given by the *pairwise disconnectivity index of a group of edges*, $F \subseteq E$, as defined in Eq. 6.

$$Dis(F) = \frac{N_0 - N_{-F}}{N_0} = 1 - \frac{N_{-F}}{N_0} \tag{6}$$

Here also, $Dis(F)$ cannot be inferred directly from the disconnectivity indices of individual edges in F.

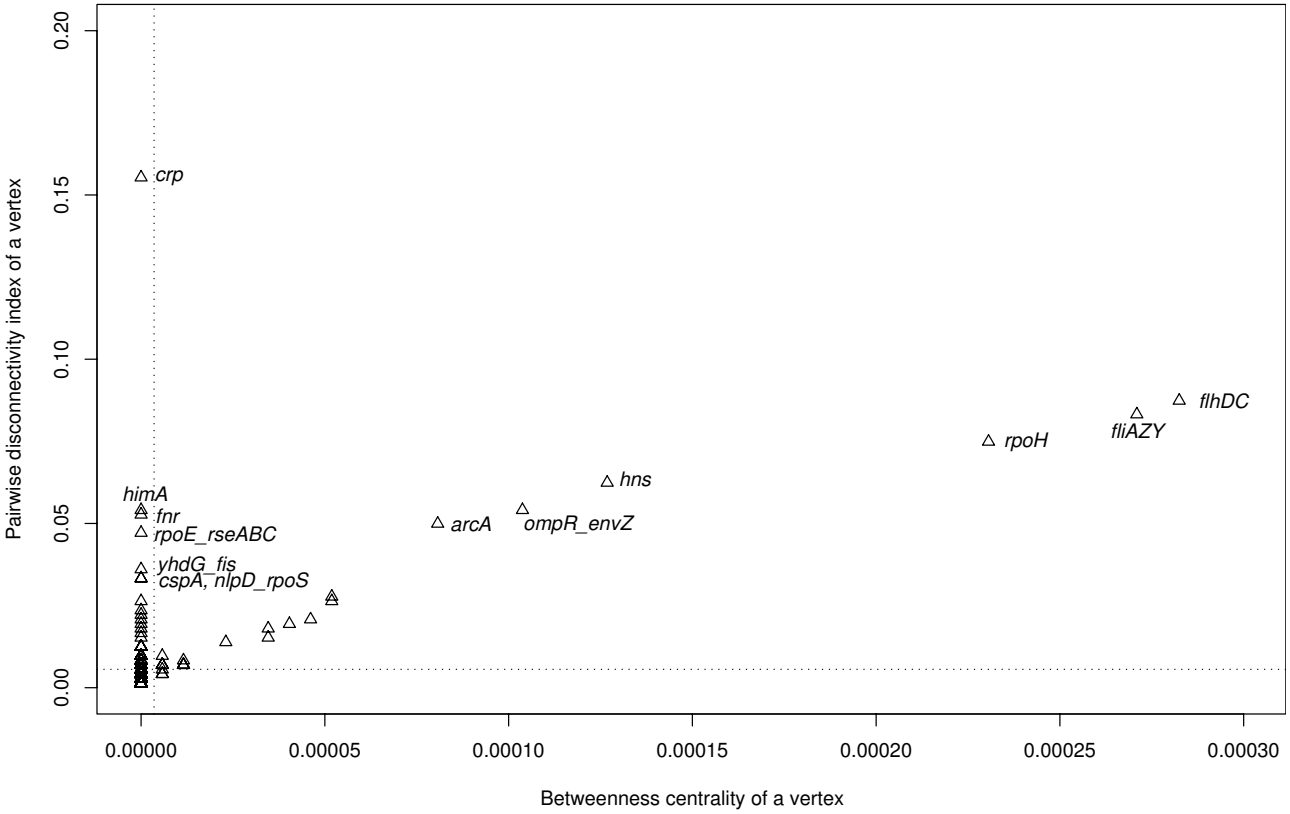


Figure 4

Betweenness centrality, *B*(*v*), and the pairwise disconnectivity index, *Dis*(*v*), of all vertices in the *E. coli* transcriptional network. The mean values of *B*(*v*) and *Dis*(*v*) are indicated with the vertical and horizontal dotted lines, respectively. Note that small values of *B*(*v*) and *Dis*(*v*) are attributed to most vertices in the network. The number of vertices in the network significantly exceeds the number of points in the plot: that is, many vertices having the same properties are represented by one point.

Applying the pairwise disconnectivity index to the analysis of biological regulatory networks

In a topological analysis of several biological networks (one signal transduction network, two transcription regulation networks, and a neuronal connectivity network), we comparatively evaluated the pairwise disconnectivity index of the individual vertices with their betweenness centrality.

Transcription networks are displayed here as directed graphs, in which the nodes represent transcription factor genes and edges represent regulatory relationships between them, i.e., the transcriptional regulation of another transcription factor gene. We used the two best characterized transcription regulation networks from organisms of different kingdoms: a bacterium (*Escherichia coli*) [20] and a unicellular eukaryote (the yeast *Saccharomyces cerevisiae*) [21].

The *E. coli* transcriptional regulatory network consists of 423 vertices and 578 edges [20]. Small values of both $B(v)$ and $Dis(v)$ are attributed to most vertices in these networks, as it can be seen from the mean values of $B(v)$ and $Dis(v)$ (Figure 4). There is a strong positive correlation between the pairwise disconnectivity indices, $Dis(v)$, and the corresponding values of betweenness centrality, $B(v)$, for many genes, among them *arcA*, *ompR_envZ*, *hns*, *rpoH*, *fliAZY*, and *flhDC*. Their $Dis(v)$ tends to be directly proportional to $B(v)$ (Figure 4). However, we have found many exceptions to this trend. These are genes that exhibit low betweenness but relatively high disconnectivity: *crp*, *himA*, *fnr*, *rpoE_rseABC*, *yhdG_fis*, *cspA*, and *nipd_rpoS*. Gene *crp* shows the highest pairwise disconnectivity index. In the network analyzed, most of these genes display both nonzero incoming degree ($k_{in} > 0$) and nonzero outgoing degree ($k_{out} > 0$) and therefore have an internal position in the network. The protein product of gene *crp* is a well-

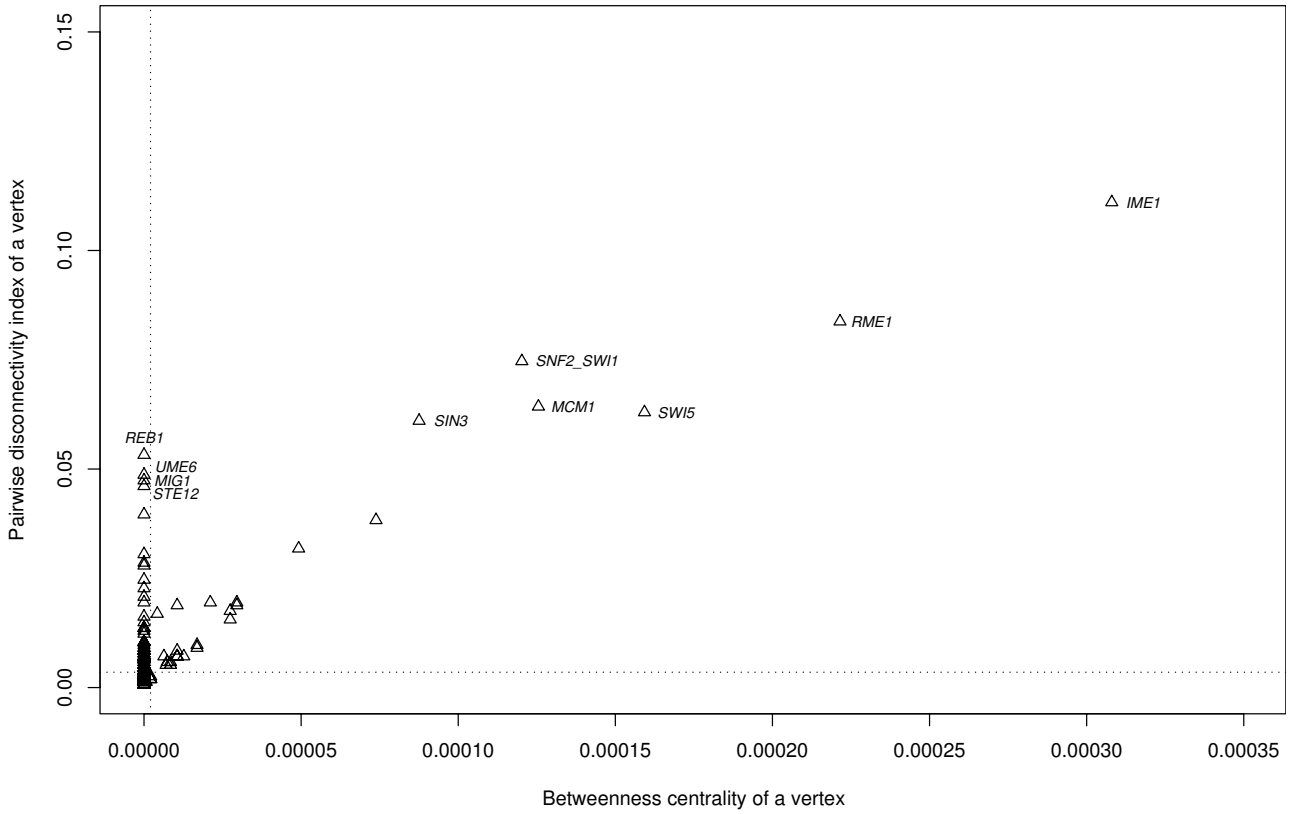


Figure 5
Distribution of betweenness centrality, *B*(*v*), and the pairwise disconnectivity index, *Dis*(*v*), in the *S. cerevisiae* transcriptional network. Small values of both *B*(*v*) and *Dis*(*v*) are attributed to most vertices, as it can be derived from the mean values of *B*(*v*) and *Dis*(*v*) (denoted with vertical and horizontal dotted lines, respectively).

characterized transcription activator triggered by cAMP and is responsible for regulating the expression of more than 100 genes in *E. coli* [22]. Moreover, genes *crp* (CRP), *fnr* (FNR) and *fis* (FIS) belong to the few global transcriptional regulators which are sufficient for directly modulating the expression of 51% of all genes in *E. coli* [23]. Betweenness centrality fails to identify them as topologically significant ones.

Similar 'predictive weakness' of betweenness centrality is observed in the transcriptional network of *S. cerevisiae* (Figure 5). This network consists of 688 vertices and 1079 edges. Again, there is a strong positive correlation between the pairwise disconnectivity index of individual genes and the corresponding value of beweenness centrality. Such genes show a diagonal positioning on the plot. Small values of both $B(v)$ and $Dis(v)$ are attributed to most vertices in these networks, which thereby exhibit low topological significance. However, many genes with $B(v) = 0$, like *REB1*, *UME6*, *MIG1*, *STE12*, have high values of $Dis(v)$ (Figure 5). In the network analyzed, all these genes exhibit no incoming degree ($k_{in} = 0$) and are therefore positioned at the periphery of the network. The relatively large value of the pairwise disconnectivity index for these genes is in accordance with the roles they play in yeast. The product of gene *REB1* (RNA polymerase I enhancer binding protein) is a DNA-binding protein that recognizes sites in both the enhancer and the promoter of rRNA transcription, as well as upstream of many genes transcribed by RNA polymerase II [24]. *REB1* is essential for cell growth: its deletion mutant is inviable [25]. The other three genes of this group (*UME6*, *MIG1*, *STE12*) have important functions too, and deleting them solicits altered phenotypes, but is not lethal [26-31] [see Additional file 1]. Among those that have equally high values of the pairwise disconnectivity index and betweeness centrality, *MCM1* is vital for the yeast cell [25,32]. Thus, at least one essential gene (*REB1*) was detected by the pairwise disconnectivity

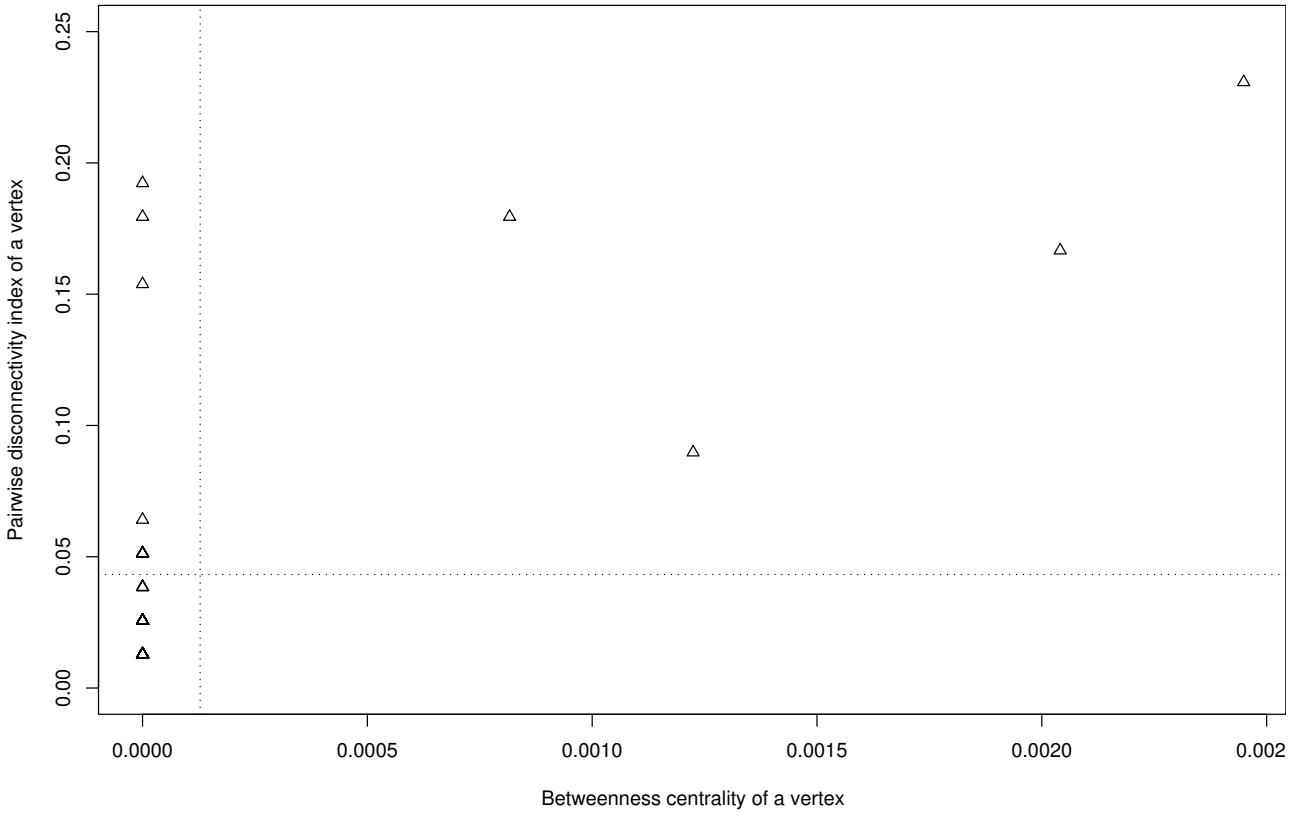


Figure 6
Comparision of betweenness centrality, *B*(*v*), and the pairwise disconnectivity index, *Dis*(*v*), of individual vertices in the neuronal connectivity network of *C. elegans*. Vertical and horizontal dotted lines stand for the mean values of *B*(*v*) and *Dis*(*v*), respectively.

index, but this gene would have been missed by betweenness centrality because of its peripheral position in the network considered.

We next analyzed the neuronal connectivity network of a simple multicellular organism, i.e. the nematode *Caenorhabditis elegans* [33]. Here, nodes represent neurons, and edges denote synaptic connections between the neurons. Each synaptic connection propagates a nerve impulse in one direction. This regulatory network includes 252 vertices and 509 directed edges. We found the same trend as in the transcription regulatory networks mentioned above: there are many vertices that display a low betweenness centrality combined with a high pairwise disconnectivity index (Figure 6): In contrast to the pairwise disconnectivity index, the betweenness centrality seems to underestimate the topological significance of some nodes, although we cannot comment here on their biological relevance since this is not documented.

The last example of regulatory networks refers to higher eukaryotes and is represented by the mammalian Toll-like receptor 4 (TLR4) signaling network. It controls a protective response of a host cell to a bacterial intervention and is important in activating the innate immunity [34,35]. The network consists of all signaling molecules that are reachable from the TLR4 receptor or from which the TLR4 receptor is reachable according to the contents of the TRANSPATH® database on signal transduction [36]. It comprises of 742 vertices (molecules) and 1952 edges (reactions) and represents a genome-wide view at a level above the individual mammalian species. The contribution of individual vertices to sustaining the integrity of these paths varies significantly with the mean pairwise disconnectivity index of 0.0044 (Figure 7). That is, an average vertex is a crucial part of only 0.44% of the existing directed paths in the TLR4 network, thereby indicating the robust topological organization of the network. There are many molecules, like Myt1 (myelin transcription fac-

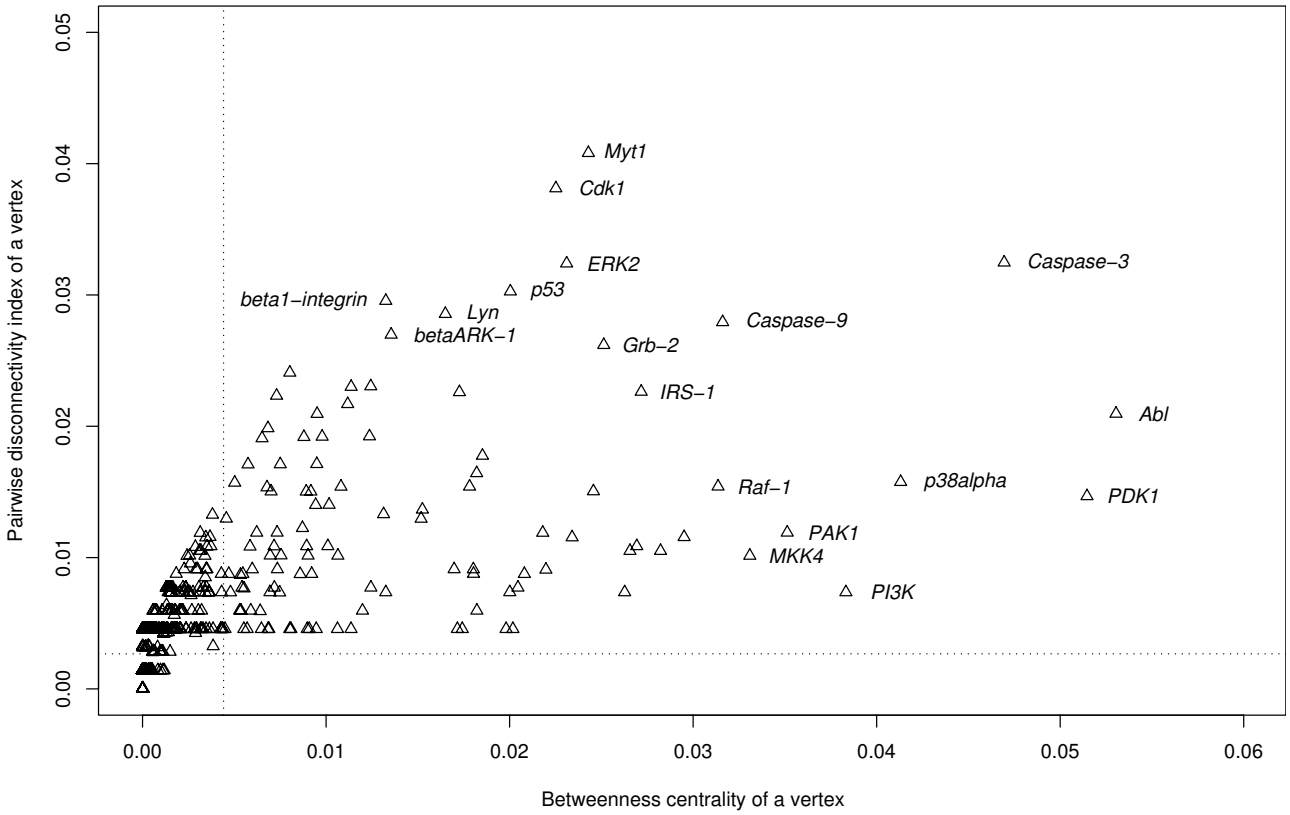


Figure 7
Plot of betweenness centrality, *B*(*v*), and the pairwise disconnectivity index, *Dis*(*v*), for each vertex in the mammalian Toll-like receptor 4 signaling network. Mean values of *B*(*v*) and *Dis*(*v*) are drawn by vertical and horizontal dotted lines at a time. Note that one point in the plot may represent many vertices having the same *B*(*v*) and *Dis*(*v*) properties.

tor 1), Cdk1 (cyclin-dependent kinase 1), ERK2 (mitogen-activated protein kinase 2), p53 (tumor suppressor p53) and others, whose disconnectivity potential significantly exceeds this average level (Figure 7). Interestingly, all of them exhibit a lethal knockout effect in mice [see Additional file 1]. The pairwise disconnectivity index of vertices positively correlates with the corresponding values of betweenness centrality. In contrast to the transcriptional regulatory networks from *E. coli* and *S. cerevisiae* and the neuron connectivity network from *C. elegans* (Figures 4, 5, 6), the mammalian TLR4 network does have vertices which exhibit both low $B(v)$ and high $Dis(v)$ values. Moreover, the relationship of the pairwise disconnectivity index and betweenness centrality in the network is much more scattered. The bigger $B(v)$ and $Dis(v)$, the broader the scattering. Thus, there are many molecules which do not differ in their $B(v)$ value, but significantly differ in their $Dis(v)$ values and *vice versa*. Molecules Abl and PDK1 display the highest levels of $B(v)$, but they are moderate in terms of $Dis(v)$. That is, Abl and PDK1 are highly engaged in shortest-path communication in the network, but there are longer paths able to sustain the communication if either Abl or PDK1 is absent. In contrast to that, molecules Myt1, Cdk1 and ERK2 show the highest values of $Dis(v)$, but they are moderate in terms of $B(v)$ which means that although these proteins are not the most significant mediators of shortest-path communication in the TLR4 network they nevertheless provide the biggest impact on the topology of the network. Altogether, all these examples demonstrate that $Dis(v)$ and $B(v)$ represent different aspects of network organization.

In order to determine the most significant vertices that are conveying the communication between others, we calculated the mediative disconnectivity indices of all vertices, $MDis(v)$, in the above mentioned networks and plotted them versus the corresponding values of betweenness centrality. The transcriptional networks from *E. coli* and *S. cerevisiae* and the neuron connectivity network from *C. elegans* show almost an ideal linear interdependence of

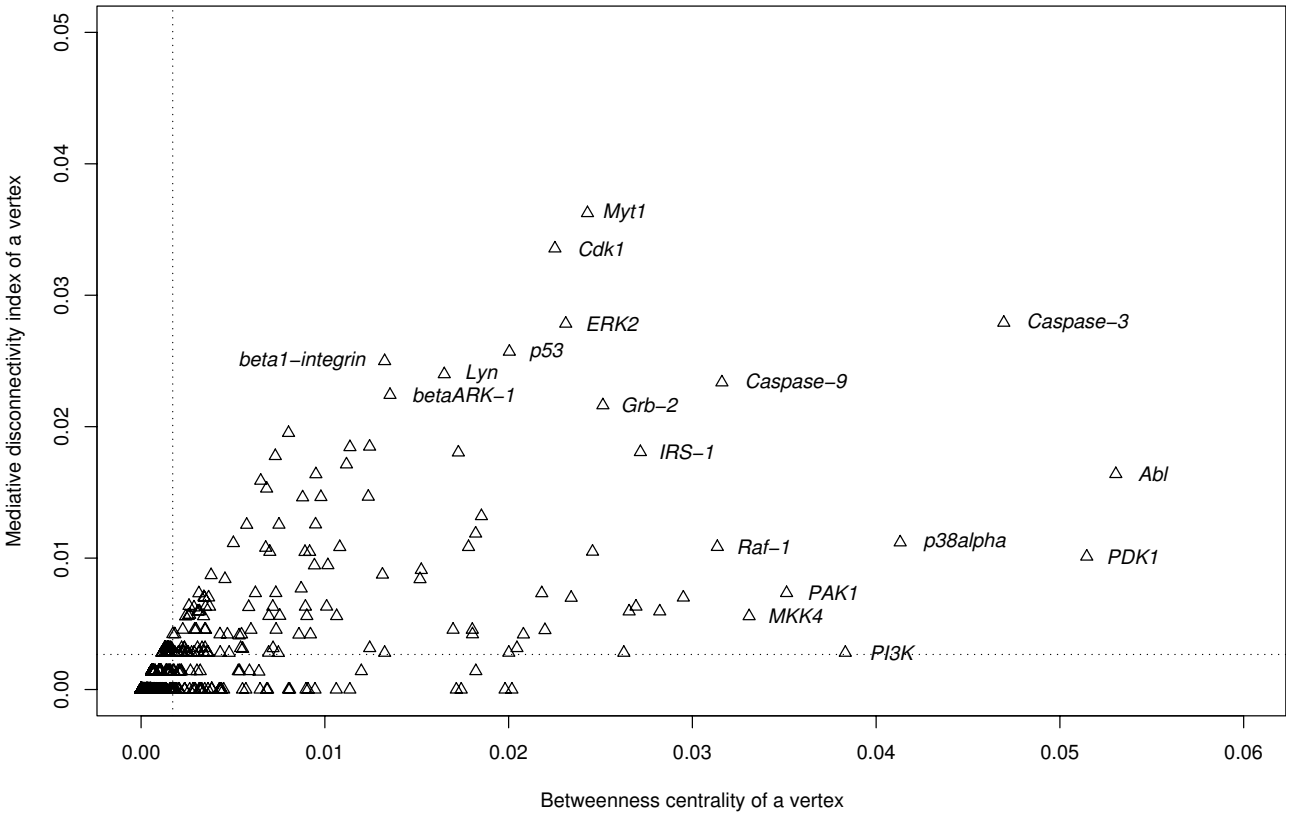


Figure 8
Relationship of betweenness centrality, *B*(*v*), and the mediative disconnectivity index, *MDis*(*v*), of individual vertices in the mammalian TLR4 network. The mean values of *B*(*v*) and *MDis*(*v*) are indicated with the vertical and horizontal dotted lines, respectively. Note that small values of *B*(*v*) and *MDis*(*v*) are attributed to most vertices in these networks.

MDis(*v*) and *B*(*v*) characterized by the correlation coefficients 0.99, 0.99 and 1.0, respectively [see Additional files 2, 3 and 4]. The corresponding mean values of *MDis*(*v*) are very small: 0.0008, 0.0006, and 0.004, respectively. Therefore, a small fraction of vertices are crucial as mediators of communication in these networks. Taken together, these networks, according to the present state of knowledge, appear to avoid significant parallelism of their paths and are relatively simply organized. In sharp contrast to that, the relationship of *MDis*(*v*) and *B*(*v*) in the mammalian TLR4 network is very scattered (Figure 8) and comparable with that of *Dis*(*v*) and *B*(*v*) (Figure 7). This network exhibits a higher complexity as compared to the previous ones. In that case, again, *MDis*(*v*) and *B*(*v*) characterize different aspects of network organization.

Discussion

Robustness is a fundamental feature of complex evolvable systems and a ubiquitously observed property of biological systems [37,38]. Robustness means the maintenance of specific functionalities of the system, i.e., its homeostasis, against perturbations and it often requires the system to change its mode of operation in a flexible way [37,38]. That can be provided at the levels of the system structure, i.e., its topology, and/or the kinetics of multiple flows between its different parts. The structural reorganization via adding or removing of vertices and edges in the network plays the primary role and is decisive. Once established, these connections may be subject to fine-tuning by modulation of the corresponding reaction kinetics. We focused here on the topological aspects of regulation.

An extremely high topological robustness can be observed in a complete graph in which each vertex has direct links to all other vertices. In a complete graph with *V* vertices, all of them have the same, maximum possible degree *V-1*. Therefore, removing a vertex or an edge provides a minimal impact on the relationships in the remaining part of the network. Such an extreme robustness excludes any flexibility and does not satisfy the multiple functional tasks of bio-

logical regulation. That might be the reason for the fact that all studied biological networks are rather sparse [1-6]: that is, the density of their edges is very low.

Highly optimized tolerance (HOT) was recently introduced as a conceptual framework to link complexity to robustness as a trade-off of kind "robust yet fragile" [39-41]. By applying similar logic to the case of regulatory networks, we propose that the topology of regulatory networks must be evolutionarily adapted to optimally combine the necessary tolerance to noisy fluctuations (both internal and external) with the necessary sensitivity to some particular inputs. In other words, the design of regulatory networks must combine robust constructions, which sustain the homeostasis of a cell and an organism, with many different flexible constructions which may allow reorganization in response to particular inputs. Intracellular regulation is basically performed via varying the sets of molecules. Identification of the basic topology of a regulatory network and its associated trade-offs is essential for understanding the role of each particular element in regulation, as well as their faults and possible countermeasures – diseases and therapies, respectively.

The topological robustness closely relates to the number of alternative (i.e., parallel) paths in a regulatory system. Here, we introduced the pairwise disconnectivity index of a network's element to characterize how crucial it is in sustaining the communication within a network. This approach can be applied to the topological analysis of a regulatory network without making any preliminary simplifications like giving preferences to shortest paths, as it is made by the betweenness centrality approach. Shortest paths represent a small fraction of all paths in a network and even the notion of shortest paths in regulatory networks is questionable because of the 'fuzzy' semantics of edges in the corresponding graphs. This fuzziness is due to the typically undefined complexity of causal relationships between network elements. A causal link from gene a to gene b, that is displayed in a network by a single edge $\{a, b\}$ and therefore appears to have length one, actually represents many steps at the level of transcription, RNA processing and splicing, transportation, translation, posttranslational modification, complex formation and so on. Thus, two edges can differ greatly in their elementary details. As a result, the path length is not a reliable variable for the analysis of such networks. The value of betweenness centrality of a given element, calculated on the basis of shortest paths passing this element, highly depends on the level of abstraction applied. Despite the very clear and attractive formalism of betweenness centrality [12,13], the practical usefulness of this measure in regard to cellular regulatory networks meets some problems due to the peculiarities of these networks.

To overcome the above mentioned shortcomings of betweenness centrality in regard to regulatory networks, all paths in the networks must be considered which is not feasible. Here, we introduced another strategy based on the fact that upon the removal of a given element some previously connected ordered pairs of vertices may become disconnected, thereby reducing the communication; this can be used to quantify the requirement of the element for the proper functioning of the whole network. Our approach emphasizes just the presence or the absence of causal links between vertices and does not rely on any assumptions concerning the meaning of these links. The pairwise disconnectivity index can be seen as a measure of topological non-redundancy of regulatory paths in a given network and, thus, as a measure of sensitivity of this network to the removal of each individual element.

The approach is rather similar to how biologists experimentally test the role of a given molecule or gene in a system of interest: the gene is knocked out or the molecule is inactivated by applying a proper inhibitor and so on. Accordingly, the evaluation of the effect of removing a vertex in a static context like a graph is the counterpart to knockout experiments performed in a lab. However, such virtual knockouts might simulate, to some extent, the corresponding wet experiments. They can be performed systematically for screening all vertices and edges in a network – which is not similarly efficiently feasible by experimental approaches. That opens up an attractive possibility to do targeted experimental verification for those elements for which a network analyses suggested topological significance. Finally, individual or groups of elements can be chosen as well for a static analysis enabling to focus on the particular context of the corresponding experiment. Altogether that might significantly contribute to a deeper understanding of network-wide interdependencies, causal relationships, and basic functional capabilities in cellular regulatory networks.

The approach has been applied to the analysis of several regulatory networks including the mammalian signal transduction TLR4 network, transcription regulatory networks from the bacteria *E. coli* and yeast *S. cerevisiae*, and the neuronal synaptic circuitry network from the nematode *C. elegans*. Different molecules, genes and neurons in these networks display a broad spectrum of pairwise disconnectivity index values, thus exhibiting a remarkable variability of the corresponding disconnectivity potentials. The impact of an individual vertex or edge is determined by its particular position in the whole network. This may be overlooked when using betweenness centrality, thereby underestimating the topological significance of some network elements.

In the $Dis(v)$-ranking of TLR4 network components (Figure 7), at least 3 out of the 4 top-ranking proteins (Cdk1, ERK2

and p53) are known as key signaling and transcription regulators in mammalian cells. All ten top-ranking genes (*Myt1*, *Cdk1*, *Caspase3*, *ERK2*, *p53*, *beta1-integrin*, *Lyn*, *Caspase9*, *betaARK-1 and Grb2*) are shown to be vital for living and developing of a mammalian organism: knockout of any of these genes causes a mutant phenotype 'inviable' [see Additional file 1]. This may serve as a benchmark that evidences the power of our method in identifying the biologically relevant key elements in regulatory networks.

By analyzing the interplay of $Dis(v)$ and $B(v)$, as well as $MDis(v)$ and $B(v)$, we have found notable difference in the organization of the mammalian TLR4 network as compared with the transcription networks from *E. coli* and yeast *S. cerevisiae*, and the neuronal synaptic network from *C. elegans* (Figures 4, 5, 6, 7, 8). The architecture of the TLR4 network exhibits a higher complexity. This might be due to various reasons: 1) the higher evolutional position of mammalian organisms, 2) the complexity of their intercellular organization, 3) differences in the organization of transcription and signal transduction networks which are adapted to different functional tasks, and 4) different completeness of our knowledge about these systems. To clarify the significance and the role of these reasons, new studies and additional analyses are necessary.

Conclusion

A new topological metric, the pairwise disconnectivity index, has been proposed. The biological importance of the suggested approach relies on its capacity to quantitatively evaluate the topological significance of each element (i.e., vertex, edge, their groups and combinations) in the context of all other elements in a given regulatory network: that is how a given network can be regulated by means of its reorganization, i.e., removing an element and restoring the element. The approach enriches the set of tools available for the analysis of biological regulatory systems.

By applying the notion of the pairwise disconnectivity index to the analysis of several regulatory networks, we show that betweenness centrality and pairwise disconnectivitiy index represent different aspects of topological organization of regulatory networks. In general, there is a positive correlation between these approaches while evaluating the topological significance of individual elements in such networks. Nevertheless, in many cases the predictive power of betweenness centrality is really poor and is not biologically relevant. The pairwise disconnectivity index provides a much broader representation of topological peculiarities of individual elements in regulatory networks.

Methods

Network databases

Literature-based databases of experimentally verified direct relationships for *Escherichia coli* [20] and *Saccharomyces cerevisiae* [21] have been used where *E. coli* is available at [42] and *S. cerevisiae* at [43]. The neuronal synaptic circuitry network of *C. elegans* was obtained from the connectivity data for *Caenorhabditis elegans* available at [33]. The mammalian TLR4 network was retrieved from the contents of the TRANSPATH® Professional database (release 7.3) on signal transduction [36] by searching for all elements that might be involved in communication with TLR4 receptor. That is, the network consists of all vertices that are reachable from TLR4 or from which TLR4 is reachable. In this network, molecules are represented at the level of "ortholog abstraction", at which all species-specific data that refer to mammalian molecules have been summarized to corresponding generic entries. Regulatory relationships between molecules and genes are displayed as semantic reactions of kind $X \rightarrow Y$ where X and Y represent signal donors and acceptors, respectively. Such a semantic style is commonly used in the literature when describing regulatory pathways. The network [see Additional files 5 and 6] can be downloaded from [44].

Selected nodes in the yeast transcriptional and the TLR4 signaling network were checked for their viability using the BIOBASE Knowledge Library™ (BKL 1.2; BIOBASE GmbH, Wolfenbuettel, Germany) and the *Saccharomyces* Genome Database (Stanford Genomic Resources [45]).

Graph analysis

Betweenness centrality

The values of betweenness centrality of vertices were computed by means of the network analysis software Pajek [46] as:

$$B(v) = \frac{1}{(n-1)(n-2)} \sum_{s \neq t \neq v \in V} \frac{\delta_{st}(v)}{\delta_{st}} \qquad (8)$$

Here, δ_{st} is the total number of shortest paths between the nodes s and t, $\delta_{st}(v)$ is those of them that pass through vertex v, and n is the number of vertices in the network. Note that the above definition represents the normalized betweenness centrality.

Pairwise disconnectivity index

The main idea of the pairwise disconnectivity index of a vertex, edge, for a group of vertices or edges is to compare the number of ordered pairs of vertices that are reachable in a graph before and after removing a vertex, edge and so on. Therefore, counting the number of ordered pairs is the essential part of any approach to determine the pairwise disconnectivity index. Various algorithms might be used for this purpose as for example depth-first (breadth-first)

search or Dijkstra's shortest paths algorithm. For the analyses described here, we have developed a tool that uses a modified depth-first search to efficiently calculate the pairwise disconnectivity indices. To estimate the index for a vertex or edge, the implemented algorithm does not exceed $\Theta(V^2)$. The program available at [47].

Statistical analysis

Besides the already mentioned software, parts of the statistical analysis have been accomplished with support of the R project for statistical computing [48].

Authors' contributions

APP developed the concept of pairwise disconnectivity indices, conceived of the study, analyzed and interpreted the data and drafted the manuscript. BG carried out the programming, performed the statistical analysis and drafted the manuscript. EW participated in the coordination of the study, helped to draft the manuscript and gave final approval of the version to be published. All authors read and approved the final manuscript.

Additional material

Additional file 1

Data on gene knockouts and their biological effects for the Dis(v)*-top-ranking elements in the networks of* E. coli, *yeast and mammalian TLR4.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-227-S1.pdf]

Additional file 2

Relationship of betweenness centrality, B(v)*, and the mediative disconnectivity index,* MDis(v)*, of individual vertices in* E. coli.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-227-S2.eps]

Additional file 3

Relationship of betweenness centrality, B(v)*, and the mediative disconnectivity index,* MDis(v)*, of individual vertices* S. cerevisiae.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-227-S3.eps]

Additional file 4

Relationship of betweenness centrality, B(v)*, and the mediative disconnectivity index,* MDis(v)*, of individual vertices in* C. elegans.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-227-S4.eps]

Additional file 5

The contents of the mammalian TLR4 network.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-227-S5.txt]

Additional file 6

The contents of the mammalian TLR4 network.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-227-S6.txt]

Acknowledgements

We are grateful to N. Voss for the methodological support in extracting network elements from TRANSPATH® database. We thank the anonymous reviewers for their critical comments and recommendations. This work has been supported in part by grant 031U110A (Intergenomics) of the German Federal Ministry of Education and Research (BMBF) and by grant 503568 (COMBIO) within the 6th Framework Programme for Research, Technological Development and Demonstration of the European Commission.

References

1. Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5:**101-113.
2. Albert R, Jeong H, Barabási AL: **Lethality and centrality in protein networks.** *Nature* 1999, **401:**130-131.
3. Albert R, Jeong H, Barabási AL: **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-382.
4. Dorogovtsev SN, Mendes JFF: **Evolution of networks.** *Adv Phys* 2002, **51:**1079-1187.
5. Albert R: **Scale-free networks in cell biology.** *J Cell Sci* 2005, **118:**4947-4957.
6. Newman MEJ: **The structure and function of complex networks.** *SIAM Review* 2003, **45:**167-256.
7. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393:**440-442.
8. Bray D: **Molecular networks: the top-down view.** *Science* 2003, **301:**1864-1865.
9. Freeman LC: **A set of measures of centrality based on betweenness.** *Sociometry* 1977, **40:**35-41.
10. Koschützki D, Schreiber F: **Comparison of Centralities for Biological Networks.** In *Proceedings of the German Conference on Bioinformatics (GCB 2004)* LNI P-53, Springer-Verlag; 2004:199-206.
11. Junker BH, Koschützki D, Schreiber F: **Exploration of biological network centralities with CentiBiN.** *BMC Bioinformatics* 2006, **7:**219-225.
12. Girvan M, Newman ME: **Community structure in social and biological networks.** *Proc Natl Acad Sci USA* 2002, **99:**7821-7826.
13. Goh KI, Oh E, Jeong H, Kahng B, Kim D: **Classification of scale-free networks.** *Proc Natl Acad Sci USA* 2002, **99:**12583-12588.
14. de Nooy W, Mrvar A, Batagelj V: *Exploratory Social Network Analysis with Pajek* Cambridge, University Press; 2005.
15. Liu WC, Lin WH, Davis AJ, Jordán F, Yang HT, Hwang MJ: **A network perspective on the topological importance of enzymes and their phylogenetic conservation.** *BMC Bioinformatics* 2007, **8:**121.
16. Potapov AP, Voss N, Sasse N, Wingender E: **Topology of mammalian transcription networks.** *Genome Inform* 2005, **16(2):**270-278.
17. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M: **The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.** *PLoS Comput Biol* 2007, **3:**e59.
18. Holme P, Huss M, Jeong H: **Subnetwork hierarchies of biochemical pathways.** *Bioinformatics* 2003, **19:**532-538.
19. Gross J, Yellen J: *Graph Theory and Its Applications* Boca Raton, CRC Press; 1998.
20. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional network of Escherichia coli.** *Nat Genet* 2002, **31:**64-68.
21. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network Motifs: Simple Building Blocks of Complex Networks.** *Science* 2002, **298:**824-827.

22. Kolb A, Busby S, Buc H, Garges S, Adhya S: **Transcriptional regulation by cAMP and its receptor protein.** *Annu Rev Biochem* 1993, **62:**749-795.
23. Martínez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria.** *Curr Opin Microbiol* 2003, **6:**482-489.
24. Morrow BE, Johnson SP, Warner JR: **Proteins that bind to the yeast rDNA enhancer.** *J Biol Chem* 1989, **264(15):**9061-9068.
25. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M: **Functional profiling of the Saccharomyces cerevisiae genome.** *Nature* 2002, **418:**387-391.
26. Strich R, Surosky RT, Steber C, Dubois E, Messenguy F, Esposito RE: **UME6 is a key regulator of nitrogen repression and meiotic development.** *Genes Dev* 1994, **8:**796-810.
27. Steber CM, Esposito RE: **UME6 is a central component of a developmental regulatory switch controlling meiosis-specific gene expression.** *Proc Natl Acad Sci USA* 1995, **92:**12490-12494.
28. Schuller HJ: **Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae.** *Curr Genet* 2003, **43(3):**139-60.
29. Carlson M: **Glucose repression in yeast.** *Curr Opin Microbiol* 1999, **2:**202-207.
30. Errede B, Ammerer G: **STE12, a protein involved in cell-type-specific transcription and signal transduction in yeast, is part of protein-DNA complexes.** *Genes Dev* 1989, **3:**1349-1361.
31. Smolka MB, Albuquerque CP, Chen SH, Zhou H: **Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases.** *Proc Natl Acad Sci USA* 2007, **104:**10364-10369.
32. Althoefer H, Schleiffer A, Wassmann K, Nordheim A, Ammerer G: **Mcm1 is required to coordinate G2-specific transcription in Saccharomyces cerevisiae.** *Mol Cell Biol* 1995, **15(11):**5917-5928.
33. ***C. elegans* – neuronal synaptic circuitry network** [http://www.wormatlas.org/neurons.htm/neuronalconninfo.htm]
34. Akira S, Takeda K: **Toll-like receptor signalling.** *Nat Rev Immunol* 2004, **4:**499-511.
35. West AP, Koblansky AA, Ghosh S: **Recognition and Signaling by Toll-Like Receptors.** *Annu Rev Cell Dev Biol* 2006, **22:**409-437.
36. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kroneberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O, Wingender E: **TRANSPATH®: An information resource for storing and visualizing signaling pathways and their pathological aberrations.** *Nucleic Acids Res* 2006, **34:**D546-D551.
37. Kitano H: **Biological robustness.** *Nat Rev Genet* 2004, **5:**826-837.
38. Hartwell L: **Robust interactions.** *Science* 2004, **303:**774-775.
39. Doyle J, Carlson JM: **Highly optimized tolerance: Robustness and design in complex systems.** *Phys Rev Lett* 2000, **84:**2529-2532.
40. Carlson JM, Doyle J: **Complexity and robustness.** *Proc Natl Acad Sci USA* 2002, **99(Suppl 1):**2538-2545.
41. Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J: **Robustness of cellular functions.** *Cell* 2004, **118:**675-685.
42. ***E. coli* transcription networks** [http://www.weizmann.ac.il/mcb/UriAlon/Network_motifs_in_coli/ColiNet-1.1]
43. **Yeast transcription network** [http://www.weizmann.ac.il/mcb/UriAlon/Papers/networkMotifs/yeastData.mat]
44. **Department of Bioinformatics. Supplementary material** [http://www.bioinf.med.uni-goettingen.de/publications/suppl_material]
45. ***Saccharomyces* Genome Database** [http://www.yeastgenome.org]
46. **Networks/Pajek. Program for Large Network Analysis** [http://vlado.fmf.uni-lj.si/pub/networks/pajek]
47. **DiVa. Program for evaluating pairwise disconnectivity indices** [http://www.bioinf.med.uni-goettingen.de/services/]
48. **The R project for statistical computing** [http://www.r-project.org]

Anhang B

Artikel 2 aus dem Jahr 2009 in BMC Systems Biology

BMC Systems Biology

BioMed Central

Methodology article

Open Access

An approach to evaluate the topological significance of motifs and other patterns in regulatory networks

Björn Goemann[1], Edgar Wingender[1,2] and Anatolij P Potapov*[1]

Address: [1]Department of Bioinformatics, Medical School, Georg August University of Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany and [2]BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbüttel, Germany

Email: Björn Goemann - bjoern.goemann@bioinf.med.uni-goettingen.de; Edgar Wingender - e.wingender@med.uni-goettingen.de; Anatolij P Potapov* - anatolij.potapov@bioinf.med.uni-goettingen.de

* Corresponding author

Published: 19 May 2009

BMC Systems Biology 2009, **3**:53 doi:10.1186/1752-0509-3-53

Received: 28 October 2008

Accepted: 19 May 2009

This article is available from: http://www.biomedcentral.com/1752-0509/3/53

Abstract

Background: The identification of network motifs as statistically over-represented topological patterns has become one of the most promising topics in the analysis of complex networks. The main focus is commonly made on how they operate by means of their internal organization. Yet, their contribution to a network's global architecture is poorly understood. However, this requires switching from the abstract view of a topological pattern to the level of its instances. Here, we show how a recently proposed metric, the pairwise disconnectivity index, can be adapted to survey if and which kind of topological patterns and their instances are most important for sustaining the connectivity within a network.

Results: The pairwise disconnectivity index of a pattern instance quantifies the dependency of the pairwise connections between vertices in a network on the presence of this pattern instance. Thereby, it particularly considers how the coherence between the unique constituents of a pattern instance relates to the rest of a network. We have applied the method exemplarily to the analysis of 3-vertex topological pattern instances in the transcription networks of a bacteria (*E. coli*), a unicellular eukaryote (*S. cerevisiae*) and higher eukaryotes (human, mouse, rat). We found that in these networks only very few pattern instances break lots of the pairwise connections between vertices upon the removal of an instance. Among them network motifs do not prevail. Rather, those patterns that are shared by the three networks exhibit a conspicuously enhanced pairwise disconnectivity index. Additionally, these are often located in close vicinity to each other or are even overlapping, since only a small number of genes are repeatedly present in most of them. Moreover, evidence has gathered that the importance of these pattern instances is due to synergistic rather than merely additive effects between their constituents.

Conclusion: A new method has been proposed that enables to evaluate the topological significance of various connected patterns in a regulatory network. Applying this method onto transcriptional networks of three largely distinct organisms we could prove that it is highly suitable to identify most important pattern instances, but that neither motifs nor any pattern in general appear to play a particularly important role *per se*. From the results obtained so far, we conclude that the pairwise disconnectivity index will most likely prove useful as well in identifying other (higher-order) pattern instances in transcriptional and other networks.

Background

Network analysis is increasingly recognized as a powerful approach to understand the organization of intracellular systems. The topology (i.e., the architecture) of a network describes how its elements are interconnected to one another, thereby providing the necessary structural basis for the subsequent analysis of the dynamics of the system. Various biological networks, such as metabolic or protein interaction networks, share global statistical features, i.e., (*i*) the small-world property referring to the shortest paths between any two vertices and highly clustered connections and (*ii*) the scale-free property, indicating that the vertex degrees follow a power-law distribution [1-7]. This implies a certain hierarchy of connectedness, as most vertices have a low degree and few vertices (hubs) have a markedly increased number of immediate neighbors.

This hierarchy is reflected in the modular organization of biological regulatory systems with each module performing its special functional task, separable from the functions of other modules [8,9]. Such a modularity of networks can be characterized topologically whereby their scale-free organization coincides with hierarchical modularity [3]. These hierarchical networks comprise many small clusters that are densely interconnected rather than consisting of independent groups of vertices [10]. Accordingly, modules may overlap with each other so that a nested type of organization is possible with smaller modules being part of bigger ones. It has been observed for various biological networks that the clustering coefficient of the vertices is approximately inversely proportional to their degree, which has been understood as the most important indication of hierarchical modularity of a network [3,11-13]. Understanding the organization of modules and their structural and functional roles emerges as a new challenge when studying biological networks. The corresponding analyses require proceeding from the level of vertices and their edges to the level of groups of these elements. It has been shown that 'network motifs' are an important feature of biological networks and may represent the simplest building blocks from which the bigger functional modules and whole networks are made [8,14,15]. They appear to relate to the lowest level of a hierarchical modularity.

Network motifs depict distinct topological patterns that occur more often in a given network than in random networks with the same size and degree distribution [14,15]. In contrast, significantly underrepresented patterns are known as anti-motifs [16]. Proteins belonging to specific motifs in the yeast protein interaction network tend to be highly conserved across species during evolution thereby underpinning that also their respective motifs may have an important, evolutionarily selected biological function [17,18]. The same network motifs have been found in diverse organisms from bacteria and yeast to plants and animals reviewed in [20]. The concept of network motifs as *the building blocks of evolution* has become one of the central topics in the analysis of complex networks. Usually, studies focus on how each network motif can carry out particular information-processing functions by means of its specific internal organization [19-23].

So far little attention has been paid to the role of motifs within a whole network, i.e., how they are embedded and how important they are for supporting the global architecture. Motifs are not isolated entities, but they are integral parts of the whole network. Thus, the targeted removal of the links among the vertices of all feed-forward loops and bi-fan clusters from the transcription regulatory network of *E. coli* fragmented this network into many small, isolated subgraphs [18]. Although this observation already indicates that motifs may be of big importance for the structure of a whole network it hides the impact of a single feed-forward loop or bi-fan representative in *E. coli*. It is unclear whether such a fragmentation is caused by a limited number of these representatives only and if the significance of a representative goes along with a particular kind of motif like the feed-forward loop. Furthermore, networks contain other topological patterns than motifs and it remains to be seen whether they take a minor role for the topology of a network [19]. Therefore, further studies are necessary and they require switching from the abstract view of a topological pattern to the level of their various representatives, the *instances of a pattern*. In general, a topological pattern depicts a unique kind of organization between a defined number of vertices which is given by the edges between these vertices. A pattern *instance* refers to a distinct set of vertices and all edges between them so that the arrangement of the edges reflects the respective pattern. To estimate the significance of such an instance for the topology of a whole network one has to consider how it relates to the rest of the network, i.e., its environment, and therewith which kind of influence it may have. No practical methods and theoretical approaches are yet available for this purpose.

To evaluate the topological significance of individual components in complex biological systems, we have recently introduced a new topological parameter – the pairwise disconnectivity index of a network's element [24]. Such an element might be a vertex (i.e., molecule, gene), an edge (i.e., reaction, interaction), as well as a group of vertices and/or edges. The pairwise disconnectivity index quantifies how essential an element is for sustaining the communication ability between all connected ordered pairs of vertices in a network. It can be viewed as a measure of sensitivity (robustness) of this network to the presence (absence) of each element. Here, we show how this concept can be used to estimate the topological

significance of a pattern instance and to find out the role of the corresponding pattern within a whole network. Subsequently, we apply this approach exemplarily to the analysis of 3-vertex topological patterns in transcription networks from different organisms: a bacterium (*E. coli*), a unicellular eukaryote (*S. cerevisiae*) and higher eukaryotes (mammals, mainly human, mouse, and rat).

Results

The topological significance of a pattern instance

Let $G = (V, E)$ be a directed graph without multiple edges that represents a regulatory network, where the vertices $v \in V$ denote biological entities, e.g., proteins, genes or small molecules. Causal relationships between these entities are made up of directed edges $e \in E$. A topological pattern is given by n connected vertices and the way they are connected with each other. The particular coherence which is described by a pattern is always based on all edges that exist between n vertices. The entirety of all distinct n-vertex patterns in G is then given by $P^n = \{P_1^n, P_2^n, \ldots, P_i^n\}$ where P_i^n is the i-th pattern consisting of n vertices. Actually, each pattern P_i^n represents a set of isomorphic connected subgraphs which have the same structural properties and differ only in the participating vertices. Accordingly, a pattern P_i^n comprehends a set of instances, i.e., $P_i^n = \{P_{i,1}^n, P_{i,2}^n, \ldots, P_{i,j}^n\}$, and each instance is a unique subgraph $P_{i,j}^n(V_{i,j}^n, E_{i,j}^n)$ of G, with the subset of vertices $V_{i,j}^n \subseteq V$ and the subset of edges $E_{i,j}^n \subseteq E$ (Figure 1). The edges in $E_{i,j}^n$ are only incident to vertices in $V_{i,j}^n$ and we denote them as the *intrinsic* edges of the pattern instance $P_{i,j}^n$. Other edges, $e \in E \backslash E_{i,j}^n$, do not contribute to the coherence of the vertices $V_{i,j}^n$. Moreover, these *extrinsic* edges are part of the environment of $P_{i,j}^n$ which describes how the pattern instance is embedded into the network. If there are more pattern instances $P_{i,j}^n$ in G than in similar random networks, then the respective pattern P_i^n is called a motif. Consequently, the entirety of n-vertex patterns in G may contain several n-vertex motifs $M^n \subseteq P^n = \{M_1^n, M_2^n, \ldots, M_i^n\}$. Then, the motif M_i^n compasses its own representatives, the instances of the motif $M_i^n = \{M_{i,1}^n, M_{i,2}^n, \ldots, M_{i,j}^n\}$.

Following the logic of [24], we denote the *topological significance of a pattern instance* $P_{i,j}^n$ as how essential for all connections within a network it is. To quantify this significance we eliminate all edges of a pattern instance (i.e., its intrinsic edges $E_{i,j}^n$) and measure how this affects the number of connected ordered pairs of vertices in the network. An ordered pair of vertices $(i, j) | i \neq j$ and $i, j \in V$, is connected iff there is at least one path from vertex i to vertex j in G. Note, that the ordered pair (i, j) is different from (j, i) in a directed network. The more ordered pairs become disconnected upon the removal of all edges of a pattern instance, the higher is the topological significance of this instance for the whole network. We define the *pairwise disconnectivity index of a pattern instance*, $Dis(P_{i,j}^n)$, as the fraction of those initially connected pairs of vertices in a network which become disconnected if the intrinsic edges of the pattern instance $P_{i,j}^n$ are removed from the network

$$Dis(P_{i,j}^n) = 1 - \frac{N'}{N}$$

In Eq. 1 N is the total number of ordered pairs of vertices in a graph $G = (V, E)$ that are connected by at least one directed path of any length. It is supposed that $N > 0$, i.e., there exists at least one edge in the network that links two different vertices. N' is the number of ordered pairs of vertices in the subgraph $G' = (V, E')$ of G where $E' = E / E_{i,j}^n$. Therefore, G' is the subgraph of G that results from removing the intrinsic edges of the pattern instance $P_{i,j}^n$ from G. The pairwise disconnectivity index of a pattern instance ranges between 0 and 1, whereas zero indicates that the removal of its intrinsic edges does not disconnect vertices within the network and one denotes the cases when no pair of vertices is connected any more.

Figure 2 illustrates how an instance of the feed-forward loop (FFL), one of the best studied network motifs [14,15,20-23], may affect the existing communication in a network. The FFL is a three-vertex pattern that is given here by the intrinsic edges $X \rightarrow Y$, $Y \rightarrow Z$ and $X \rightarrow Z$. It is linked to the rest of a network by its vertices X, Y, Z where each of these can be at the start or end of an extrinsic edge (blue dotted edges). Further extrinsic edges are between other pairs of vertices in the environment of a FFL instance, e.g. the ordered pair (E_1, E_2). Whether a FFL instance can have an impact on the connection between

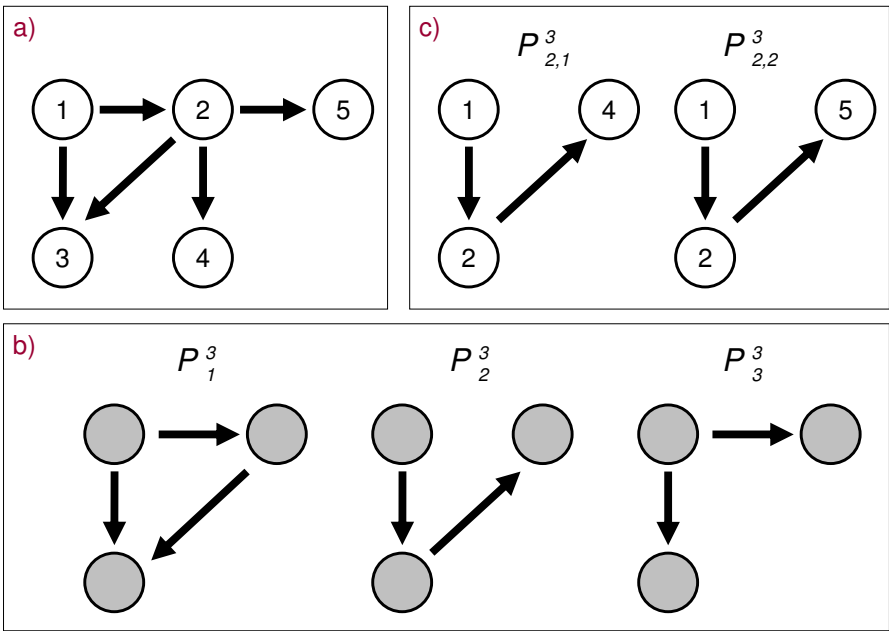


Figure 1

Topological patterns and their instances in a network. **A**: A toy network with the set of vertices $V = \{1,..., 5\}$. **B**: The entirety of all 3-vertex patterns $P^3 = \left\{P_1^3, P_2^3, P_3^3\right\}$ in the toy network. **C**: The two instances of the pattern P_2^3.

two vertices depends on the kind of constituents of the paths that link them. If these paths consist of extrinsic edges only then the connection will not be affected upon the removal of the FFL (e.g., the pair (E_1, E_3)). Essentially, the FFL instance may be critical for those paths which include at least one intrinsic edge of the instance. However, that depends on the presence of alternative (i.e., parallel) paths between the corresponding vertices that use extrinsic edges only. For example, the pair (E_2, Y) does not critically depend on the FFL instance due to the presence of another path (E_2, X, E_3, E_4, Y) that includes no intrinsic edge of the instance. In contrast, the pair (E_2, E_6) looses its connection upon the deletion of the FFL instance though three parallel paths are connecting these vertices.

Usually, several instances of a particular pattern can be found in a network. For estimating the topological significance of the pattern itself the impact of its representatives has to be considered. We find that the average pairwise disconnectivity index of all instances of a pattern reflects this appropriately and define

$$Dis(P_i^n) = \overline{Dis(P_{i,j}^n)} = \frac{1}{J}\sum_{j=1}^{J} Dis(P_{i,j}^n)$$

as the *pairwise disconnectivity index of a pattern* P_i^n that consists of J instances. With it Eq. 2 also states the topological significance of a randomly chosen instance of the pattern P_i^n.

Applying the pairwise disconnectivity index to the analysis of topological patterns in regulatory networks

We have applied our approach to the characterization of three-vertex topological patterns in transcription regulation networks from three different organisms: a bacteria (*Escherichia coli*) [14], a unicellular eukaryote (the yeast *Saccharomyces cerevisiae*) [15] and higher eukaryotes (mammals: human, mouse, rat) [25,26]. 3-vertex motifs were identified by means of the Z-Score as proposed by Alon and colleagues [15]. This normalized value states whether the abundance of a pattern in the real network exceeds its occurrence in a number of random ensembles: that is, a positive Z-Score refers to an over-representation in the real network, whereas a negative Z-Score means under-representation. Since there is no commonly accepted threshold Z-Score value for defining motifs, we consider patterns with Z-Score > 0 as motifs and all other ones as non-motifs. For the networks of *E. coli* and *S. cerevisiae* 3-vertex motifs were already identified [14,15], whereas for the mammalian transcription network this is reported for the first time. To distinguish between different motifs many of which have no commonly accepted names, we used the identification numbers (IDs) of small connected graphs as it is provided by the FANMOD software [27,28]. The name of a pattern instance was gener-

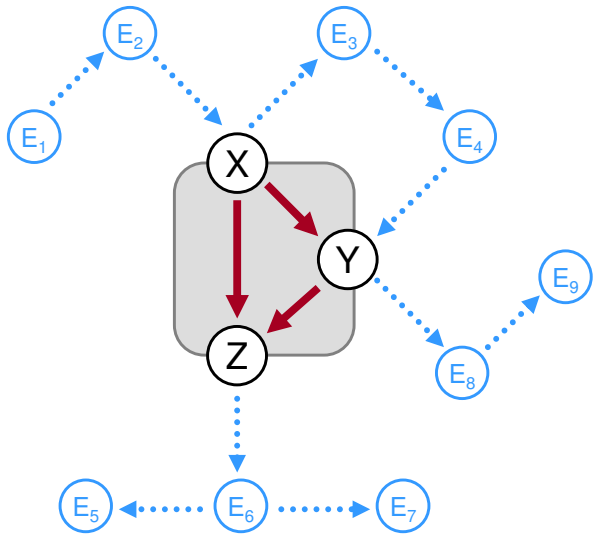


Figure 2
The embedding of a feed-forward loop (FFL) instance into a network. The FFL pattern is given here by the coherence between the vertices *X*, *Y*, *Z*. Therewith its only instance consists of these vertices and the intrinsic edges $X \rightarrow Y$, $Y \rightarrow Z$, $X \rightarrow Z$. How the FFL instance relates to the rest of the network is determined by those kinds of extrinsic edges that are attached to the vertices *X*, *Y*, *Z* (blue dotted edges). Other extrinsic edges link further vertices (blue vertices) in the environment of the FFL instances. The connection between a pair of vertices can be affected only then by the FFL instance if the paths linking them contain at least one of the intrinsic edges. For example, the connection between the pair (E_2, E_6) depends of the relation between the vertices *X*, *Y*, *Z*. In contrast, there still is an alternative path between the vertices E_2 and *Y* that remains untouched. Note that the 'feed-forwarding' action of the FFL instance does not apply to those paths which cross only one intrinsic edge of this instance – e.g., path {E_1, E_2, *X*, *Y*, E_8, E_9} and path {E_3, E_4, *Y*, *Z*, E_6, E_7}.

ated by combining a prefix *E*, *Y* or *M* for referring to *E. coli*, *S. cerevisiae* or mammalian, respectively, with the corresponding ID followed by the pairwise disconnectivity index rank of the instance among all instances of a given pattern.

Bacterial transcription network

The *E. coli* transcription network consists of 418 vertices and 519 edges. It exhibits four 3-vertex patterns, two of which are motifs according to the Z-score criteria (Figure 3). One of these motifs (ID = 6) appears most frequently and seems to be part of larger motifs known as the single-input module [20]. The mean pairwise disconnectivity index of its instances is 0.0039: that is only about 0.4% of all connected pairs of genes become suspended when a randomly selected instance of this motif is deleted from the network. The second motif, ID = 38, is known as the feed-forward loop [14,15] and appears in the *E. coli* network less often than the previous, but its instances exhibit a higher average pairwise disconnectivity index (0.018). The patterns ID = 12 and ID = 36 are not over-represented here (negative Z-Score) and are therefore not ranked as motifs. The pattern ID = 12 denotes a chain-like structure where a gene regulates another one which itself regulates a third one. It is attributed to a pairwise disconnectivity index that ranges within the same scale as the feed-forward loop on average. In contrast, the pattern ID = 36, that abstracts the influence of two genes on a third one, has a much lower mean pairwise disconnectivity index than that of the ID = 12 pattern, but higher than that of the ID = 6 motif.

The boxplots in Figure 4 show how the pairwise disconnectivity index is distributed among the instances of different 3-vertex patterns (see Figure 3, *E. coli*). The population of each pattern is very heterogeneous. Most instances exhibit a low pairwise disconnectivity index value. However, very few pattern instances cause a significant effect when deleted, thereby indicating that the network is vulnerable against a targeted removal of particular instances. While about 3% of all motif instances are not crucial for sustaining the connection between any gene pair, nearly 9% of them disconnect at least 1% of the gene pairs in *E. coli*. In contrast, the instances of non-over-represented patterns always disconnect at least one gene pair and one third of them 1% or more. In general, comparing the medians of pattern instances (shown in Figure 4 as a solid horizontal bar) indicates that motifs are not topologically more significant than the non-motif patterns.

		E. coli			*S. cerevisiae*			*Mammals*		
Pattern	*ID*	*Freq*	*Z-Score*	$\overline{Dis}$	*Freq*	*Z-Score*	$\overline{Dis}$	*Freq*	*Z-Score*	$\overline{Dis}$
	6	4777	11.23	0.0039	11892	14.54	0.0018	1916	-0.39	0.0023
	12	160	-11.21	0.0189	295	-14.25	0.0135	1068	-1.67	0.011
	14	-	-	-	18	-1.30	0.0063	73	-10.47	0.0079
	36	227	-11.76	0.0046	894	-13.86	0.0019	1620	-2.91	0.0044
	38	42	11.18	0.018	70	14.00	0.0086	129	5.68	0.0050
	46	-	-	-	-	-	-	17	11.18	0.0042
	78	-	-	-	-	-	-	4	-10.85	0.0099
	102	-	-	-	1	15.91	0.028	3	0.35	0.0224
	140	-	-	-	-	-	-	1	-0.88	0.0137
	164	-	-	-	-	-	-	197	-6.52	0.0051
	166	-	-	-	1	4.65	0.0052	20	7.12	0.0058
	174	-	-	-	-	-	-	6	7.31	0.0109
	238	-	-	-	-	-	-	1	-	0.0073

Figure 3

3-vertex patterns in the transcriptional networks of *E. coli*, *S. cerevisiae* and *Mammals*. The column *Pattern* outlines the respective pattern. Its name can be found in the column *ID*. The column *Freq* denotes the number of occurrences of a pattern (a positive *Z-Score* indicates over-representation). $\overline{Dis}$ stands for the mean pairwise disconnectivity index of all instances of a pattern.

Nevertheless, the instance with the highest pairwise disconnectivity index in the *E. coli* network is a motif instance. This feed-forward loop consists of the genes *hns*, *flhDC* and *fliAZY* (Figure 5, ID = E.38.1). Interestingly, the gene *flhDC* is part of all pattern instances with a high topological significance, either together with the gene *fliAZY* or *ompR_envZ* (Figure 5). Like *hns* and *fliAZY*, the gene *flhDC* is involved in the synthesis of flagella in *E. coli*. A reduced activity of *flhDC* and *fliAZY* results in the loss of motility in *E. coli* [29,30] which has vital consequences for the bacteria. This can be the case for a loss of the *ompR_envZ* regulatory system too, which is known to play a critical role in stress response by regulating the transcription of porin genes in response to medium osmolarity [31]. Altogether, the high topological significance of the pattern instances in Figure 5 seems to reflect the importance of the few recurring interactions between these essential genes for *E. coli* adequately.

Yeast transcription network

The transcription network of *S. cerevisiae* consists of 688 vertices and 1079 edges. It features three additional patterns besides those ones that have already been identified in *E. coli*. A positive Z-Score is attributed to four patterns in *S. cerevisiae*, although the patterns ID = 102 and ID = 166 occur only once (Figure 3). Likewise to the observations from *E. coli*, the average topological significance of the motif ID = 6 is lower than that of the feed-forward loop. On average, a randomly selected FFL instance breaks the connection between less than 1% of all connected pairs of genes, which is lower than for instances of the pattern ID = 12. Their mean pairwise disconnectivity index is about 0.0135 and appears to be the highest of all patterns in the *S. cerevisiae* network with a negative Z-Score.

Except for the pattern ID = 14, the pairwise disconnectivity index varies considerably for the instances of a pattern

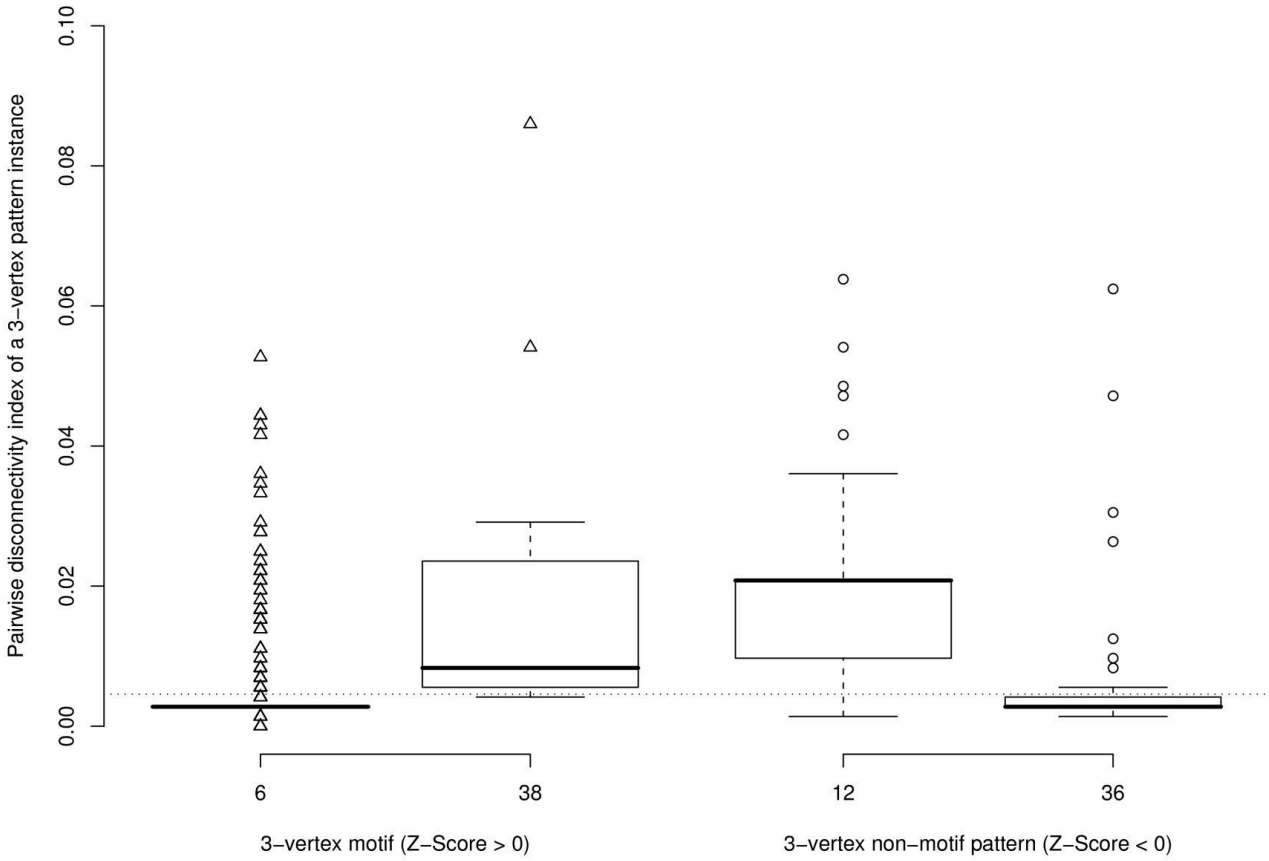


Figure 4

The topological significance of 3-vertex pattern instances in *E. coli*. The left boxplot denotes the two 3-vertex motifs found in the network on the x-axis and the distribution of the pairwise disconnectivity index of their instances on the y-axis. The right boxplot constitutes this for patterns that are not over-represented in *E. coli*. The dotted line indicates the average pairwise disconnectivity index of all pattern instances in the network (0.0046). Note that one point may stand for several pattern instances.

in this network (Figure 6). The respective patterns of the candidates with a high topological significance display positive Z-Scores as well as negative Z-Scores, which refer to over-representation and under-representation, respectively. Hence, motifs are not in favour for sustaining the pairwise connections between genes compared with non-motif patterns. In contrast to the *E. coli* network, the *S. cerevisiae* network seems to be more robust upon the elimination of a pattern instance, since much less of them have a notable effect on the existing pairwise connections between genes at all: The average pairwise disconnectivity index of a pattern instance is with 0.002 just half as high as in the *E. coli* network. Therewith, more alternative paths are at hand that strengthen pairwise connections between genes here so that also fewer instances cause a significant perturbation in the network (about 3% with $Dis\ (P^3_{i,j}) \geq$ 0.05 in yeast contrary to 10% in the *E. coli* network). Certainly, the overall impact of these pattern instances is comparable to the *E. coli* network (see Figures 5 and 7). A reason for this might be that such pattern instances are embedded in an alike fashion in both networks and may so have a similar influence on the existing connections.

The highest pairwise disconnectivity index is about 0.08 (Figure 6) and refers to a feed-forward loop instance that embodies the genes *RME1*, *IME1* and *IME1_UME6* (Figure 7). *RME1* is known to encode a zinc finger protein that can repress the transcription of *IME1* [32]. *RME1* and *IME1* are the master regulators of meiosis in *S. cerevisiae* [33-35]. An *ime1* disruption prevents expression of almost all meiotic genes and all tested meiotic events [33]. *RME1* is essential for sustaining the communication abilities between lots of gene pairs, similar to the genes *MCM1*, *SNF2_SWI1* and *SWI5*. Gene *MCM1* is central to the tran-

Pattern	*ID*	*Dis*	*Participants (X, Y, Z)*
	E.38.1	0.0859	hns, flhDC, fliAZY
	E.12.1	0.0638	ompR_envZ, flhDC, fliAZY
	E.36.1	0.0624	crp, ompR_envZ, flhDC
	E.12.2	0.0541	crp, flhDC, fliAZY
	E.38.2	0.0541	flhDC, fliAZY, fliLMNOPQR
	E.38.3	0.0541	flhDC, fliAZY, fliFGHIJK
	E.38.4	0.0541	flhDC, fliAZY, fliE
	E.38.5	0.0541	flhDC, fliAZY, flhBAE
	E.38.6	0.0541	flhDC, fliAZY, flgBCDEFGHIJK
	E.6.1	0.0527	flhDC, fliAZY, flgAMN
	E.6.2	0.0443	ompR_envZ, flhDC, csgDEFG
	E.6.3	0.0443	ompR_envZ, flhDC, fadL

Figure 5

The highest topologically significant 3-vertex pattern instances in *E. coli*. The column *Pattern* outlines the respective pattern of an instance. Its name can be found in the column *ID*. The column *Dis* refers to the pairwise disconnectivity index of the pattern instance. The column *Participants* denotes the set of genes involved in the instance and their locations within the pattern.

scription control of cell-type specific genes and the pheromone response. The *SNF2/SWI* complex is an evolutionarily conserved ATP-dependent chromatin remodeling complex that plays an important role in DNA damage repair, DNA replication and stress response [36]. *SWI5* activates the expression of cell cycle genes [37]. Altogether, these genes exert vital functions in *S. cerevisiae* and each of them appears quite frequently among the pattern instances with the highest topological significance.

Mammalian transcription network

The third network represents genes coding for transcription factors in mammalian species (human, mouse, and rat) and their interplay. This mammalian network consists of 279 vertices and 657 edges and has been extracted from the contents of the TRANSPATH® database on signal transduction [25] and the TRANSFAC® database on eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors [26]. Unlike the other two networks it contains all of the thirteen possible 3-vertex patterns. Although five patterns display positive Z-Scores, only four of them indicate a clear over-representation (Figure 3). In addition, one might find it difficult to classify the pattern ID = 102 as a motif due to its low frequency. Nevertheless, the FFL is a motif in mammals and the only pattern that is over-represented in all three networks. Although its occurrence rises with the increasing density and complexity of the networks, its topological significance is decreasing notably. Actually, a low average pairwise disconnectivity index can be observed for almost all motifs in mammals, with motif ID = 174 as the only exception.

Three of the seven patterns with a negative Z-Score have been found in the networks of *E. coli* and *S. cerevisiae* too, but unlike in mammals the pattern ID = 6 is a motif in them. Yet, its average topological significance for these networks does not differ greatly. Similar applies to the pattern ID = 12 that exhibits one of the highest mean pairwise disconnectivity indices here as well. In contrast, just a minor role seems to be adopted by the pattern ID = 36 though it is the second most common one. Other non-motif patterns in the mammalian network are crucial for linking only 1% of gene pairs mostly on average. Nevertheless, their appearance is a hint on the more complex

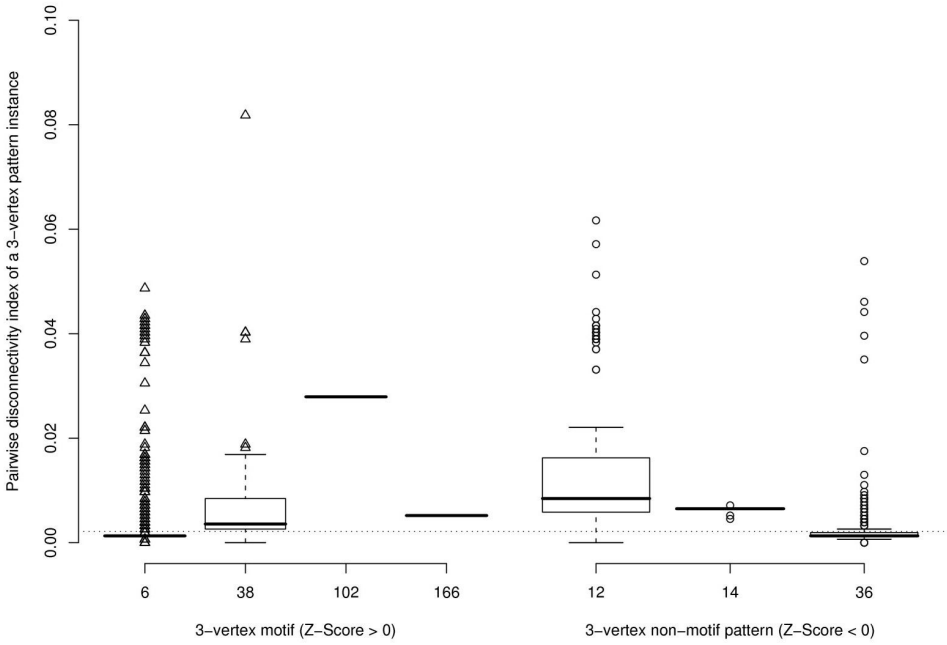


Figure 6

Distribution of the pairwise disconnectivity index in 3-vertex patterns in *S. cerevisiae*. The left boxplot shows how the pairwise disconnectivity index of a motif instance (x-axis) is distributed in within the respective motif (y-axis). The boxplot on the right present a similar comparison for non-motif patterns in *S. cerevisiae*. The overall mean pairwise disconnectivity index of all pattern instances in the network (0.0021) is represented by the dotted line. One point may represent several pattern instances.

organization of transcription regulation in higher organisms. Thus, it seems to be convenient that the pattern ID = 238 can be found only here (Figure 3): it represents the mutual transcription control of three retinoic acid receptor isoforms with the vertices *RAR-alpha*, *RAR-beta* and *RAR-gamma*. Note that this pattern does not even occur in any random network of similar size and degree distribution. On the other hand, it is still surprising that the pattern ID = 164 appears nearly 200 times in the mammalian network, but neither in the network of *E. coli* nor in the network of *S. cerevisiae*.

Despite the overall low mean topological significance of the various patterns in the mammalian network, the pairwise disconnectivity index of their instances covers a broad range of values (Figure 8). This spreading is even stronger for non over-represented patterns and more noticeable as in the other two networks. Thus, a high topological significance does not go along with motifs here as well. However, this network is different with regard to the robustness of its architecture: About one third of all pattern instances do not affect any of pairwise connections between genes and more than 15% disconnect at least 1% of the gene pairs. No motif instance exhibits a pairwise disconnectivity index higher than 0.04. This can be found for non-over-represented patterns exclusively (ID = 6, 12, 36, 164).

The most intense perturbation outranks the topologically most significant pattern instances in the other two networks. Deleting this pattern instance, which comprises the genes *c-myc, HMGA1* and *PAX3*, suspends the connections between 10% of all genes in the mammalian network (M.6.1, Figure 9). The proto-oncogene *c-myc* is engaged in diverse processes ranging from cell proliferation to apoptosis [38] and its interaction with *PAX3* repeatedly occurs in the pattern instances with the highest topological significance (Figure 9). Such a frequent appearance has been observed for some genes in the networks of *E. coli* and *S. cerevisiae* too. Furthermore, these genes have been found to exert vital functions in their organism. The same applies for *PAX3* and *c-myc* in mammals: The paired box gene 3 activates developmental genes (e.g., *Mitf*) and just as *c-myc* the loss of *PAX3* is lethal [39]. It is interesting to note that all transcription factors encoded by the genes constituting the interlinked

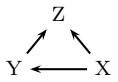
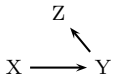
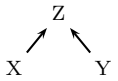
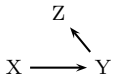
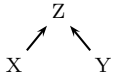
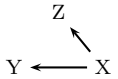
Pattern	*ID*	*Dis*	*Participants (X, Y, Z)*
	Y.38.1	0.0818	RME1, IME1_UME6, IME1
	Y.12.1 Y.12.2	0.0616 0.0571	SIN3, SNF2_SWI1, MCM1 SNF2_SWI1, MCM1, SWI5
	E.36.1	0.0538	SIN3, SWI5, RME1
	Y.12.3	0.0513	MCM1, SWI5, RME1
	Y.36.2	0.0461	MCM1, REB1, SWI5
	Y.6.1 Y.6.2 Y.6.3	0.0487 0.0435 0.0435	SIN3, RME1, SNF2_SWI1 SNF2_SWI1, ALPHA1, MCM1 SNF2_SWI1, HAP4, MCM1

Figure 7

The highest topologically significant 3-vertex pattern instances in *S. cerevisiae*. The column *Pattern* outlines the respective pattern of an instance. Its name can be found in the column *ID*. The column *Dis* refers to the pairwise disconnectivity index of the considered instance. The column *Participants* denotes the set of genes involved in the instance and their locations within the pattern.

patterns M.6.1, M.6.2, M.12.1, M.12.2 and M.164.1 (Figure 9) play pronounced roles in cell proliferation (*E2F-1*, *c-myc*, *c-fos*, *HMGA1*, and *NSEP1*) or are important developmental regulators (*PAX3*, *Mitf*).

A note on the joint deletion of intrinsic edges

The unusually often appearance of the same links (i.e., intrinsic edges) between genes in the pattern instances with the highest pairwise disconnectivity indices in all three networks raises the question of their contribution to the estimated significance of these pattern instances. Probably, the removal of individual intrinsic edges may already destroy the connection between many gene pairs so that their simultaneous removal is not as crucial. Otherwise they may have a significant non-additive impact taken together. However, answering this requires knowing the effect of deleting a single interaction (i.e., edge) in a network which can be accomplished in a similar way as for a pattern instance. It has been introduced as the pairwise disconnectivity index of an edge in [24] and specifies the fraction of ordered pairs becoming disconnected due to the removal of an individual edge.

As a first attempt, this fraction has been estimated for each intrinsic edge of a pattern instance in the three networks and their sum has been opposed to the pairwise disconnectivity index of the respective pattern instance. Although such kind of comparison highlights just a tendency if and how far the intrinsic edges of a pattern instance act synergistically, it is already a way that works for all kinds of patterns independent of their specific arrangement. Figure 10 illustrates this approximation for the pattern instances with the highest pairwise disconnectivity index in each network. The edge weights denote the topological significance of an edge for the corresponding network, e.g., $Dis(hns \rightarrow flhDC) = 0.005$ for the edge from gene *hns* to *flhDC* in *E. coli*. Hence, the deletion of this interaction merely disconnects a half percent of all pairwise linked genes in *E. coli*. As expected, no effect is accomplished by removing the edge from *hns* to *fliAZY*,

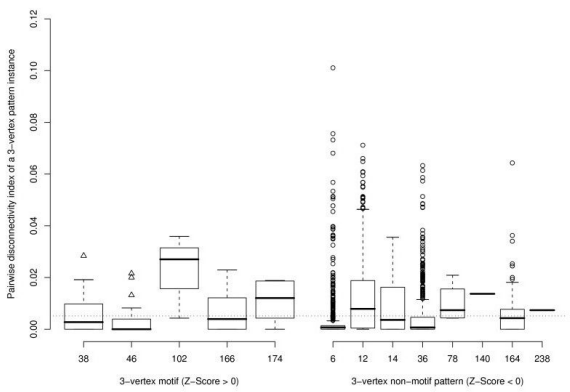


Figure 8
The pairwise disconnectivity index of 3-vertex pattern instances in *Mammals*. The boxplot on the left indicates the distribution of the pairwise disconnectivity index in the 3-vertex motifs in the network. The same relation is pictured in the right boxplot for patterns with a negative or no Z-Score at all. The dotted line describes the mean pairwise disconnectivity index of all 3-vertex pattern instances in the network (0.0051). One point may represent several pattern instances.

since there is always the alternative path via *flhDC*. In contrast, a relatively high pairwise disconnecivity index has been measured for the edge from *flhDC* to *fliAZY*. But still, the summarized effect of deleting these intrinsic edges separately from the *E. coli* network (0.049) is considerably lower as compared with the topological significance for the whole pattern instance, $Dis(E.38.1) = 0.086$. The same holds for the other two pattern instances in Figure 10 as well. Therewith a much stronger impact on pairwise connections between genes clearly exists due to the coherence of the intrinsic edges.

Whether this can be generalized for all pattern instances found in the three networks is shown in Figure 11. Most pattern instances in the three networks cluster near the diagonal since the joint removal of their intrinsic edges disconnects approximately the same number of gene pairs as the separate elimination of them does. However, some exceptions have been found, especially among those patterns that exhibit a high pairwise disconnectivity index *per se*.

A pattern instance is positioned below the diagonal dotted lines in Figure 11 due to considerable overlapping in the sets of pairwise linked genes which become disconnected upon the separate removal of the intrinsic edges of the instance. For example, consider how the vertices *1* and *5* in Figure 1A are linked. To disconnect them it is enough to delete one of the edges $1 \rightarrow 2$ or $2 \rightarrow 5$ at a time. Such kinds of dependencies seem to exist in larger scales in the analyzed networks pinpointing to lots of gene pairs that are connected in a linear chain-like manner as reflected by the pattern ID = 12 (Figure 3). There are almost no independent alternative paths between such gene pairs so that the connection between them is very sensitive upon the deletion of a single intrinsic edge. Therewith, the pattern ID = 12 is contained virtually exclusively amongst the pattern instances below the diagonal dotted lines in Figure 11.

The concurrent elimination of the intrinsic edges of a pattern instance located above the diagonal dotted lines breaks also pairwise connections between genes that are not so easily assailable as described above. At least two paths between such genes exist, each using a unique combination of intrinsic edges. Thus, they cannot be affected by eliminating a single intrinsic edge only. For example, in Figure 2 there are three paths linking vertex E_2 with E_6: The first one includes the intrinsic edge $X \rightarrow Z$. The second consists of the intrinsic edges $X \rightarrow Y$ and $Y \rightarrow Z$ whereas the third path contains only the edge $Y \rightarrow Z$. However, no matter which of the intrinsic edges is deleted, the vertex pair (E_2, E_6) remains untouched since at least one of the three paths is still present. Their connection is disrupted only if the whole pattern is deleted. Such dependencies can be observed in Figure 11 for few pattern instances in *E. coli*, but increasingly in the other two networks. This trend is most distinctive in the mammalian network. Besides the pattern instances with a high pairwise disconnectivity index, a considerable number of motif instances appear in the lower left corner of the plot for the mammalian network (Figure 11, red triangles): their intrinsic edges have an extremely small or even no impact at all on pairwise connections between genes. But as motif instances, they are a bottleneck for linking many gene pairs.

Discussion

A new method to asses the global role of patterns and motifs

The work presented here describes a method that has been proved to be suitable for evaluating the role of topological patterns within a network. This holds true regardless of the size and complexity of these patterns. The method assesses the significance of a pattern depending on the contribution of its instances, i.e. connected subgraphs, for the connectivity of a network. The approach is based on the technique described previously in [24], which estimates the necessity of a network element (e.g., a vertex or an edge) for sustaining the communication ability between connected pairs of vertices in a network. This is accomplished in a similar way as wet experiments in a lab: a gene (corresponding to a vertex in a graph) is knocked out and the effect of this removal is observed in the considered context. The same may be applied to a reaction

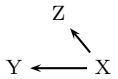
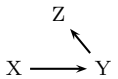
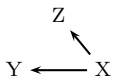
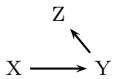
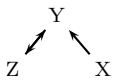
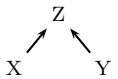
Pattern	*ID*	*Dis*	*Participants (X, Y, Z)*
	M.6.1 M.6.2 M.6.3	0.1011 0.0756 0.0731	c-myc, HMGA1, PAX3 c-myc, NSEP1, PAX3 IRF-1, FOXA2, NR3C1
	M.12.1	0.0711	E2F-1, c-myc, PAX3
	M.6.4	0.0681	NR3C1, C/EBPAlpha, NR1l3
	M.12.2	0.0660	c-myc, PAX3, Mitf
	M.164.1	0.0643	Mitf, c-fos, HMGA1
	M.36.1 M.36.2	0.063 0.061	POU1F1, RUNX2, ER1 c-myc, MYCN, PAX3

Figure 9
The highest topologically significant 3-vertex pattern instances in *Mammals*. The column *Pattern* outlines the respective pattern of an instance. Its name can be found in the column *ID*. The column *Dis* refers to the pairwise disconnectivity index of the considered instance. The column *Participants* denotes the set of genes involved in the instance and their locations within the pattern.

(an edge in the graph), when a gene has been mutated and the encoded product (vertex) is still present, but unable to undergo a certain reaction.

In this work, we have proposed to proceed likewise for pattern instances, but disturbing the interactions between the involved vertices rather than eliminating the vertices themselves. Consequently, only the causal links between these vertices are destroyed and therewith the respective pattern is removed in a minimally invasive way. This is conducted without making any *a priori* assumptions on the analyzed network and its properties. In contrast to the attempt made in [18], we destroy the coherence between the edges of only one single pattern instance at a time, leaving the remainder of the network intact. On the one hand, different impacts on the network connectivity exerted by the various instances of a pattern can thus be discovered. On the other hand, the topological role of a pattern can be determined more realistically since an overrating is avoided.

3-Vertex patterns in transcriptional networks

We exemplarily applied the method developed and proposed here to the analysis of transcriptional regulation networks of three very distinct taxa (*E. coli*, *S. cerevisiae* and mammals, i. e. human, mouse, and rat); for simplicity, we focused here on 3-vertex topological patterns in these networks, but the method can easily be adopted to the analysis more complex and larger patterns. A first check of which of the thirteen possible 3-vertex patterns are present in these networks at all revealed that all of them can be found in the mammalian network, the *S. cerevisiae* network contains seven and that of *E. coli* only four of them. Moreover, these latter four patterns are shared by all three networks. Amongst them, only the "feed-forward loop" is statistically over-represented and, thus, could be considered as a "motif" (Figure 3).

As to be expected, the abundance of a pattern decreases with its complexity: Thus, 3-vertex patterns with two edges occur much more frequently than those with three edges,

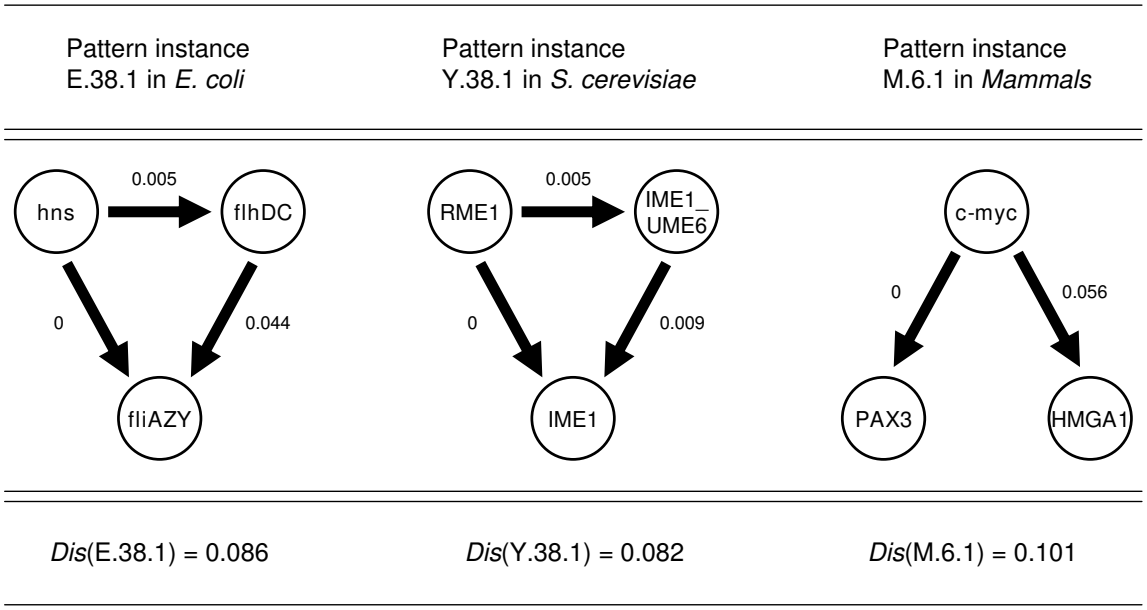


Figure 10
The impact of the coherence between the genes in the pattern instances with the highest topological significance in the transcription networks. Edge weights denote the impact of an individual edge for sustaining the pairwise connections between genes. In all three pattern instances, the intrinsic edges themselves fail to reproduce the impact as accomplished by their common elimination. The summarized effects of their separate removal are significantly lower as compared with the simultaneous deletion of the relation between the respective genes. Hence, these pattern instances affect only as whole entities those gene pairs that are linked by several alternative paths.

etc. The order of the abundance is almost the same in all three transcription networks. It is of interest that the network patterns "coupled feedback loop" (Figure 3, ID = 78) and "3-vertex-circuit" (Figure 3, ID = 140) do not exist in the networks of *E. coli* and *S. cerevisiae* and are clearly under-represented in the network of mammals (Figure 3), although they are widespread in signaling circuits of various bacterial and eukaryotic organisms [40-44]. We assume that this is an intrinsic property of transcriptional networks and cannot be explained by the incompleteness of the underlying knowledge, since other patterns of similar complexity (e.g., the mentioned feed-forward loop) are not consistently under-represented among these three networks.

All networks studied here appear to be rather robust against the elimination of a randomly chosen pattern instance. Therewith, the various 3-vertex patterns in these networks display a low topological significance on average. Mostly, the overall majority of the instances of a pattern have a rather small effect on the existing pairwise connections between genes, in most cases even less than 1% of all pairwise connections are affected.

Pattern instances can be identified that are crucial for the connectivity of the network

Motifs do not seem to be more important than non-motif patterns for the global architecture of a network

Also the motifs among the 3-vertex patterns examined did not exhibit a generally higher importance for the connectivity of the whole network than non-motif patterns, as one might have expected. This is, however, in agreement with previous studies on the evolutionary and functional assessment of motifs in the regulatory networks of different yeasts, which have provided evidence that motifs are not subject to any particular evolutionary pressure to preserve the corresponding interaction pattern [45,46]. No simple relationships have been found between evolutionary conservation and over-representation of network patterns, on the one hand, and their functional enrichment, on the other hand, in the yeast regulatory network [42]. In accordance with these observations, our results indicate that there is no positive correlation between the abundance (i.e., over-representation) of a network pattern and its topological significance. Thus, focusing on motifs exclusively rather than searching for important pattern instances in general would have lead to a completely different and deceptive picture.

In spite of the generally low impact of all types of patterns (including motifs) found in the analyzed networks, a few

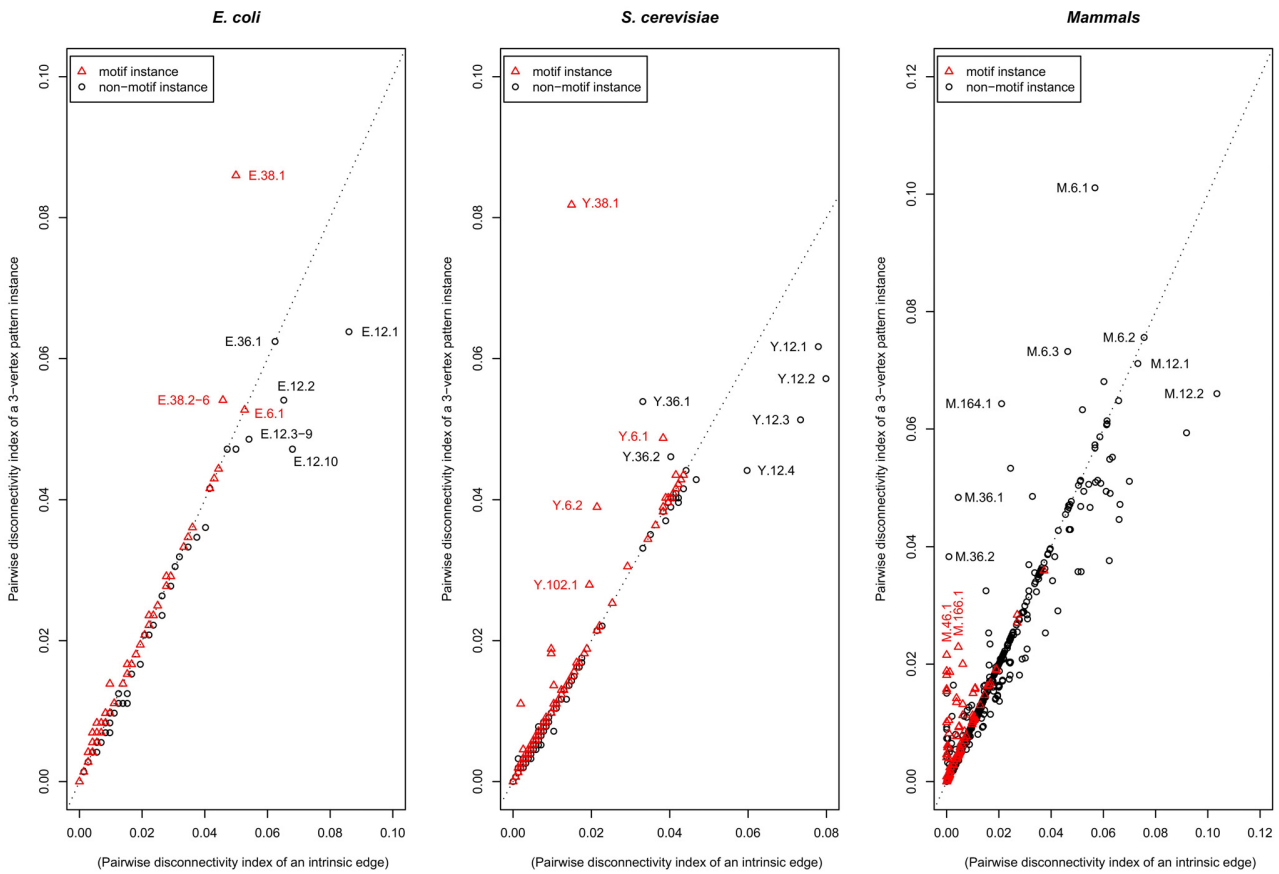


Figure 11

The joint deletion of the intrinsic edges of a pattern instance may synergistically reduce connections between genes in the *E. coli*, *S. cerevisiae* and *Mammals* networks. The pairwise disconnectivity index of a pattern instance (motifs are drawn as red triangles, others as black circles) is outlined on the y-axis. By contrast, the x-axis denotes the fraction of gene pairs becoming disconnected upon the deletion of a single intrinsic edge, summarized for all intrinsic edges of a given pattern instance. The diagonal dotted lines indicate the cases when the impact of the concurrent elimination of all intrinsic edges of a pattern instance does not differ from the sum of impacts provided by the separate removal of the same edges. Hence, pattern instances that are drawn below these lines include intrinsic edges that have an overlapping in their impacts: they can disconnect the same gene pairs. Finally, the position of a pattern instance above the dotted lines shows that some of its intrinsic edges are parts of alternative (i.e., parallel) paths between two genes. Such genes do not become disconnected when only one intrinsic edge is eliminated, but some of them do upon simultaneous removal of all intrinsic edges of the pattern instance: i.e., the joint removal exerts a higher than merely additive effect.

pattern instances cause a significant perturbation upon their removal. This trend is manifested in the heterogeneous distribution of the pairwise disconnectivity index among all the instances of a pattern (Figs. 3, 4, 5). Topologically, this may originate from the way how a pattern instance is embedded, i.e., its particular position within the whole context of the respective network. Biologically, such heterogeneity might be caused by the influence of the genes in the network that are forming a pattern instance. In the networks, the topologically most significant pattern instances consist preferably of genes that provide basic functions for the organism. Interestingly, most of these instances belong to one of the patterns that are shared by the three networks, which may emphasize the importance of these patterns. Furthermore, such instances may indicate locations within the networks rendering them vulnerable upon a targeted removal.

Among the pattern instances that are of particular importance for the network connectivity, motif instances again do not play a predominant role over instances from non-over-represented patterns. In the mammalian network, most of the outliers even belong to the non-motif patterns. Altogether, our data support the view that far not all instances of any pattern (motif or not), but only few of them may play specific functional roles [47] and thereby

exhibit a strong impact on pairwise connections between genes in transcription networks.

Pattern instances of high topological significance tend to form clusters

In all the networks analyzed here, a limited number of genes repeatedly appears in the pattern instances displaying the highest topological significance. For example, in *E. coli* the gene *flhDC* is part of all pattern instances that disconnect at least 4% of the gene pairs, preferably together with the genes *fliAZY* or *ompR_envZ*. Similar observations can be made in *S. cerevisiae* for the genes *MCM1*, *SIN3*, *SNF2_SWI1* and *SWI5*. Likewise in the mammalian network, the interaction between the genes *c-myc* and *PAX3* participates in many of the pattern instances with a high pairwise disconnectivity index. Altogether, the common occurrence of genes and interactions between them underlines the key importance of these constituents for the corresponding organism. All these genes are engaged in important processes and at least in *E. coli* and *S. cerevisiae* they are crucial for linking a significant number of gene pairs [24]. Hence, their damage can be lethal for the respective organism. Furthermore, these pattern instances are not located in different regions of a network. They are connected with each other and seem to form a bigger pattern cluster that controls a lot of pairwise connections between genes in these networks.

Edges of pattern instances display synergistic effects

In many cases, the intrinsic edges of a pattern instance contribute to its pairwise disconnectivity index in a synergistic manner, i.e., the simultaneous removal of the respective edges exerts a much higher than merely additive effect (Figure 11). Although the approach we used for this purpose is a conservative approximation, it shows a principal tendency in these networks. More exact computations of this feature may be desirable but developing suitable algorithms for this, which have to take into account the particular characteristics of every pattern separately, was beyond the scope of this paper. However, we find that our approach was adequate to disclose clearly that the intrinsic edges of certain pattern instances display synergistic effects. This is the case for the pattern instances with the highest pairwise disconnectivity index in each of the three networks. Some other candidates have been found in *E. coli* and increasingly more in *S. cerevisiae* and mammals. This trend goes along exactly with the increasing density of the networks (1.2 edges per vertex in the *E. coli* network, 1.6 in *S. cerevisiae* and 2.3 in *Mammals*). The reason for this is on the hand: a more densely connected network provides a higher average vertex degree and thereby offers more alternative paths between pairs of vertices. These paths need not to share a similar set of edges, i.e., the connection of a pair is becoming more robust requiring more edges to be removed in order to disconnect it.

Prospects of the proposed method

It should be noted that the observations reported here have been made for the networks as they are known at present. In particular the mammalian network may still suffer from incomplete knowledge. However, our method can be used for monitoring changes in such networks obtained from updated pathway databases like TRANSPATH® [25] in the future. We see our results as the beginning of a large work which may consider the analysis of increasingly larger patterns including more than 3 vertices. More regulatory networks of various types (e.g., signal transduction networks, protein-protein interaction networks, gene expression networks) from different organism must be considered and tested in this regard in future as well. First attempts with signaling networks have confirmed the basic conclusions drawn here in spite of small characteristic differences in some details. Thus, we feel that the basic trends reported here will hold true for the more complete transcriptional as well as for other types of networks that will come up in future with increased reliability of high-throughput approaches and their systematic application.

On the other side, our method provides for the first time the possibility to assess the impact of patterns and motifs in general as well as individual pattern instances onto the overall connectivity of a graph. It is therefore suitable to identify bottlenecks in a biological network, which may be particularly important for the normal function of a cell, and may be top candidates to investigate disease mechanisms related to these functions. Since it identifies individual components in a network (vertices, edges, or pattern instances), it works independently of any *a priori* knowledge about the statistical over- or under-representation of certain network features. Though our approach was developed for the analysis of biological regulatory networks, it seems to be suitable for the analysis of other networks regardless of the particular nature of processes they represent (e.g., ecological, social, technical networks).

Conclusion

We have developed a new method that quantifies how the elimination of a topological pattern instance affects the existing communication abilities within a network. We have applied this method exemplarily to the analysis of 3-vertex topological patterns and their instances in the transcription networks from a bacteria, yeast and mammals.

The elimination of most 3-vertex pattern instances does not drastically affect the global structure of transcription networks. However, these networks are vulnerable upon a

targeted perturbation of few pattern instances. In these cases, the links between their genes contribute to the pairwise disconnectivity index of the pattern instance in a synergistic manner, i.e., the simultaneous removal of the respective edges exerts a much higher than merely additive effect. The topological significance of an instance does not easily correlate with the abundance of the respective pattern in a network. Although motifs might play an essential role in their respective local contexts, they do not seem to be more important than non-motif patterns for the global architecture of a network. Rather, the topological role of a pattern instance is unique and mainly determined by its location and the way how it is embedded in a given network.

Methods

Network databases

Literature-based databases of experimentally verified direct relationships for *Escherichia coli* [14] and *Saccharomyces cerevisiae* [15] have been used where *E. coli* V1.1 and *S. cerevisiae* V1.3 are available at http://www.weizmann.ac.il/mcb/UriAlon. The mammalian network of transcription factor genes (human, mouse, rat) was retrieved from the TRANSPATH® Professional database (release 8.3, made in 2007) on signal transduction [25] and TRANSFAC® Professional database (release 11.3, made in 2007) on eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors [26]. The network describes the causal relationships between genes that are coding for transcription factors, based on the regulation of these genes from transcription factors. However, the transcription factors themselves are not part of the network, i.e., the interaction chain "gene A codes for transcription factor A regulates gene B" has been summarized to: "gene A → gene B", which is a commonly used technique when inferring gene regulatory networks. Furthermore, genes are represented at the level of "ortholog abstraction", at which all species-specific data (human, mouse, rat) that refer to mammalian genes have been summarized to corresponding generic entries.

Selected genes (vertices) in the yeast and mammalian transcription networks were checked for their viability using the BIOBASE Knowledge Library™ http://www.biobase.de and the *Saccharomyces* Genome Database (Stanford Genomic Resources [48]).

Pattern analysis

The networks were scanned for 3-vertex topological patterns using the FANMOD software with default settings [27,28]. The statistical significance of the network motifs was evaluated by means of the Z-Score [15], $Z = (M_{real} - M_{rand})/SD$, where M_{real} and M_{rand} are the numbers of appearance of the motif in the real network and the randomized networks, respectively. *SD* is the standard deviation. The sign of edges (such as 'positive' for activation or 'negative' for inhibition) is not considered.

The pairwise disconnectivity index was calculated using the DiVa software [49]. The statistical analysis was accomplished with R [50].

The pairwise disconnectivity index of an edge

For estimating the impact of a single intrinsic edge on the existing pairwise connections between genes we have applied the pairwise disconnectivity index on an edge as defined in [24]. In this manner it states the fraction of those ordered pairs of vertices that have been disconnected upon the removal of an edge, i.e., $Dis(e) = 1 - \frac{N'}{N}$. Similar to Eq. 1, N is the number of linked ordered pairs of vertices in a network and we assume $N > 0$. The term N' stands for the number of connected ordered pairs of vertices in the network we obtain when deleting the edge e. Hence, $Dis(e) = 0$ the edge e is not crucial for linking at least of vertex pair. In contrast, $Dis(e) = 1$ if no vertex pairs remains connected.

Authors' contributions

APP and BG conceived the study, interpreted the data and drafted the manuscript. BG carried out the programming and performed the statistical analysis. EW participated in the coordination of the study and gave final approval of the version to be published. All authors read and approved the final manuscript.

Acknowledgements

This work has been supported in part by grant 031U110A (Intergenomics) of the German Federal Ministry of Education and Research (BMBF) and by grant 503568 (COMBIO) within the 6th Framework Programme for Research, Technological Development and Demonstration of the European Commission.

References

1. Albert R, Jeong H, Barabási AL: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
2. Albert R, Jeong H, Barabási AL: **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-382.
3. Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5:**101-113.
4. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393:**440-442.
5. Dorogovtsev SN, Mendes JFF: **Evolution of networks.** *Adv Phys* 2002, **51:**1079-1187.
6. Newman MEJ: **The structure and function of complex networks.** *SIAM Review* 2003, **45:**167-256.
7. Albert R: **Scale-free networks in cell biology.** *J Cell Sci* 2005, **118:**4947-4957.
8. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402:**C47-C52.
9. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks.** *Proc Natl Acad Sci USA* 2003, **100:**12123-12128.
10. Ravasz E, Barabási AL: **Hierarchical organization in complex networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67(2 Pt 2):**026112.

11. Dorogovtsev SN, Goltsev AV, Mendes JF: **Pseudofractal scale-free web.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2002, **65(6 Pt 2):**066122.
12. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.
13. Potapov AP, Voss N, Sasse N, Wingender E: **Topology of mammalian transcription networks.** *Genome Inf Ser* 2005, **16:**270-278.
14. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional network of Escherichia coli.** *Nat Genet* 2002, **31:**64-68.
15. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: Simple building blocks of complex networks.** *Science* 2002, **298:**824-827.
16. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U: **Superfamilies of evolved and designed networks.** *Science* 2004, **5:**1538-1542.
17. Wuchty S, Oltvai ZN, Barabási AL: **Evolutionary conservation of motif constituents in the yeast protein interaction network.** *Nat Genet* 2003, **35:**176-179.
18. Dobrin R, Beg QK, Barabási AL, Oltvai ZN: **Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network.** *BMC Bioinformatics* 2004, **5:**10.
19. Vázquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabási AL: **The topological relationship between the large-scale attributes and local interaction patterns of complex networks.** *Proc Natl Acad Sci USA* 2004, **101:**17940-17945.
20. Alon U: **Network motifs: theory and experimental approaches.** *Nat Rev Genet* 2007, **8:**450-461.
21. Mangan S, Alon U: **Structure and function of the feed-forward loop network motif.** *Proc Natl Acad Sci USA* 2003, **100:**11980-11985.
22. Mangan S, Zaslaver A, Alon U: **The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.** *J Mol Biol* 2003, **334:**197-204.
23. Kalir S, Mangan S, Alon U: **A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli.** *Mol Syst Biol* 2005, **1:**2005.0006.
24. Potapov AP, Goemann B, Wingender E: **The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks.** *BMC Bioinformatics* 2008, **9:**227.
25. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kroneberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O, Wingender E: **TRANSPATH®: An information resource for storing and visualizing signaling pathways and their pathological aberrations.** *Nucleic Acids Res* 2006, **34:**D546-D551.
26. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34:**D108-D110.
27. Wernicke S, Rasche F: **FANMOD: a tool for fast network motif detection.** *Bioinformatics* 2006, **22:**1152-1153.
28. **FANMOD** [http://www.minet.uni-jena.de/~wernicke/motifs/]
29. Soutourina O, Kolb A, Krin E, Laurent-Winter C, Rimsky S, Danchin A, Bertin P: **Multiple control of flagellum biosynthesis in Escherichia coli: role of H-NS protein and the cyclic AMP-catabolite activator protein complex in transcription of the flhDC master operon.** *J Bacteriol* 1999, **181:**7500-7508.
30. Bertin P, Terao E, Lee EH, Lejeune P, Colson C, Danchin A, Collatz E: **The H-NS protein is involved in the biogenesis of flagella in Escherichia coli.** *J Bacteriol* 1994, **176:**5537-5540.
31. Pratt LA, Hsing W, Gibson KE, Silhavy TJ: **From acids to osmZ: multiple factors influence synthesis of the OmpF and OmpC porins in Escherichia coli.** *Mol Microbiol* 1996, **20:**911-917.
32. Toone WM, Johnson AL, Banks GR, Toyn JH, Stuart D, Wittenberg C, Johnston LH: **Rme1, a negative regulator of meiosis, is also a positive activator of G1 cyclin gene expression.** *EMBO J* 1995, **14:**5824-5832.
33. Mitchell AP: **Control of meiotic gene expression in Saccharomyces cerevisiae.** *Microbiol Rev* 1994, **58:**56-70.
34. Bowdish KS, Yuan HE, Mitchell AP: **Positive control of yeast meiotic genes by the negative regulator UME6.** *Mol Cell Biol* 1995, **15:**2955-2961.
35. Rubin-Bejerano I, Mandel S, Robzyk K, Kassir Y: **Induction of meiosis in Saccharomyces cerevisiae depends on conversion of the transcriptional represssor Ume6 to a positive regulator by its regulated association with the transcriptional activator Ime1.** *Mol Cell Biol* 1996, **16:**2518-2526.
36. Osley MA, Tsukuda T, Nickoloff JA: **ATP-dependent chromatin remodeling factors and DNA damage repair.** *Mutat Res* 2007, **618:**65-80.
37. McBride HJ, Yu Y, Stillman DJ: **Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation.** *J Biol Chem* 1999, **274:**21029-21036.
38. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10:**789-799.
39. Li J, Liu KC, Jin F, Lu MM, Epstein JA: **Transgenic rescue of congenital heart disease and spina bifida in Splotch mice.** *Development* 1999, **126:**2495-2503.
40. Venkatesh KV, Bhartiya S, Ruhela A: **Multiple feedback loops are key to a robust dynamic performance of tryptophan regulation in Escherichia coli.** *FEBS Lett* 2004, **63:**234-240.
41. Brandman O, Ferrell JE, Li R, Meyer T: **Interlinked fast and slow positive feedback loops drive reliable cell decisions.** *Science* 2005, **310:**496-498.
42. Ramsey SA, Smith JJ, Orrell D, Marelli M, Petersen TW, de Atauri P, Bolouri H, Aitchison JD: **Dual feedback loops in the GAL regulon suppress cellular heterogeneity in yeast.** *Nat Genet* 2006, **38:**1082-1087.
43. Kim D, Kwon YK, Cho KH: **Coupled positive and negative feedback circuits form an essential building block of cellular signaling pathways.** *BioEssays* 2007, **29:**85-90.
44. Kim JR, Yoon Y, Cho KH: **Coupled feedback loops form dynamic motifs of cellular networks.** *Biophys J* 2008, **94:**359-365.
45. Mazurie A, Bottani S, Vergassola M: **An evolutionary and functional assessment of regulatory network motifs.** *Genome Biology* 2005, **6:**R35.
46. Meshi O, Shlomi T, Ruppin E: **Evolutionary conservation and over-representation of functionally enriched network patterns in the yeast regulatory network.** *BMC Syst Biol* 2007, **1:**1.
47. Konagurthu AS, Lesk AM: **On the origin of distribution patterns of motifs in biological networks.** *BMC Systems Biology* 2008, **2:**73-81.
48. ***Saccharomyces* Genome Database** [http://www.yeastgenome.org]
49. **DiVa. Program for evaluating the pairwise disconnectivity index** [http://www.bioinf.med.uni-goettingen.de/services/]
50. **The R project for statistical computing** [http://www.r-project.org]

Anhang B

Artikel 3 aus dem Jahr 2009 in Genome Informatics Series Volume 23

COMPARATIVE ANALYSIS OF TOPOLOGICAL PATTERNS IN DIFFERENT MAMMALIAN NETWORKS

BJOERN GOEMANN[1]
bjoern.goemann@bioinf.med.uni-goettingen.de

ANATOLIJ P. POTAPOV[1]
apo@bioinf.med.uni-goettingen.de

MICHAEL ANTE[1]
michael.ante@bioinf.med.uni-goettingen.de

EDGAR WINGENDER[1,2]
ewi@bioinf.med.uni-goettingen.de

[1] *Department of Bioinformatics, University Medical Center Goettingen, Georg August University Goettingen, Goldschmidtstr. 1, D-37077 Goettingen, Germany*
[2] *BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbuettel, Germany*

We have systematically analyzed various topological patterns comprising 1, 2 or 3 nodes in the mammalian metabolic, signal transduction and transcription networks: These patterns were analyzed with regard to their frequency and statistical over-representation in each network, as well as to their topological significance for the coherence of the networks. The latter property was evaluated using the *pairwise disconnectivity index*, which we have recently introduced to quantify how critical network components are for the internal connectedness of a network. The 1-node pattern made up by a vertex with a self-loop has been found to exert particular properties in all three networks. In general, vertices with a self-loop tend to be topologically more important than other vertices. Moreover, self-loops have been found to be attached to most 2-node and 3-node patterns, thereby emphasizing a particular role of self-loop components in the architectural organization of the networks. For none of the networks, a positive correlation between the mean topological significance and the *Z*-score of a pattern could be observed. That is, in general, motifs are not *per se* more important for the overall network coherence than patterns that are not over-represented. All 2- and 3-node patterns that are over-represented and thus qualified as motifs in all three networks exhibit a loop structure. This intriguing observation can be viewed as an advantage of loop-like structures in building up the regulatory circuits of the whole cell. The transcription network has been found to differ from the other networks in that (*i*) self-loops play an even higher role, (*ii*) its binary loops are highly enriched with self-loops attached, and (*iii*) feed-back loops are not over-represented. Metabolic networks reveal some particular topological properties which may reflect the fact that metabolic paths are, to a large extent, reversible. Interestingly, some of the most important 3-node patterns of both the transcription and the signaling network can be concatenated to subnetworks comprising many genes that play a particular role in the regulation of cell proliferation.

Keywords: network topology, network motif analysis, transcription network, signaling network, metabolic network, pairwise disconnectivity index

1 Introduction

Triggered by the increasingly better defined paradigms of Systems Biology, biological systems are described most appropriately as the sum of processes they exert rather than comprehensive catalogs of objects they constitute of. It has been proven for the various biological processes that it is most feasible to represent them as networks or, more formally, as graphs, which comprise the participating components as nodes and their relations as edges. Once the architecture of a network, or its wiring diagram, is known, it

is possible to add quantitative information to the edges in order to proceed to dynamical simulations of the system's behavior under certain conditions. This is still hard to achieve for a complete system, as the holistic approach of systems biology would demand. However, it can be done for defined subsystems. In the most basic case, these are topological patterns which describe the connection between a few nodes. Amongst them, motifs are believed to be the simplest building blocks in biological networks. Consequently, they have been explored intensively in the last years and have been investigated for their behavior in diverse kinds of networks [1].

Different functions of a living cell can be represented by different networks such as the metabolic, the signaling or the gene regulatory network. The latter is mostly considered as a mere transcription network, but inclusion of post-transcriptional mechanisms as well becomes increasingly feasible [1]. Holistic modeling of a system would obviously require an integrated view of these different networks, but before that task can be tackled, we have to make ourselves aware about their particularities. It is therefore the goal of this contribution to characterize the three mentioned networks, reconstructed for mammalian cells, with regard to their topological patterns and the impact of these substructures for the whole respective networks.

For this purpose, we have proposed a new topological parameter, the *pairwise disconnectivity index* [3], which is useful in identifying network components that are most critical for the coherence of a network. The methodology may be applied on single nodes or edges, or whole subgraphs such as motifs, as evidenced for transcription networks [4].

Here, we apply this logic to topological patterns of different sizes on the mammalian metabolic, signal transduction and transcription network to investigate whether these networks differ significantly in content and impact of certain topological patterns such as self-edges and 3-vertex-patterns.

2 Methods

2.1 Construction of the Networks

The mammalian transcription network was retrieved from the TRANSFAC® database, release 11.3 [5], and the TRANSPATH® database, release 8.3 (BIOBASE, Wolfenbuettel, Germany) [6]. In this network, the nodes represent transcription factor (TF) genes, and the edges the genetic interactions between them, i.e. comprising expression of each gene and trans-activation/-repression of the target genes of its product. The TRANSPATH database was also used to reconstruct the signal transduction network for mammalian (mostly human, mouse and rat) cells; for this, we extracted "semantic" reactions only which focus on the essential components between which information is actively forwarded, and did so on the level of "orthogroups" [7]. Both networks therefore represent "reference networks", i.e. superpositions of all reactions and paths that have been identified in any mammalian species, in any tissue / cell type. The transcription

network includes 279 nodes and 658 edges, while the signaling network is made by 1571 nodes and 3425 edges.

The mammalian metabolic network was reconstructed from Ligand section of the KEGG database [8], comprising all genes encoding metabolic enzyme activity in mammalian (more precisely: human, mouse and rat) systems. To keep this network comparable with the other two, we chose a gene-centric view here as well, so that the nodes represent genes encoding metabolic enzymes, and the edge semantics is to forward a metabolite produced by one enzyme to one that consumes it. The metabolic network consists of 1793 nodes and 5538 edges.

2.2 Computation of the Pairwise Disconnectivity Index

In a directed graph $G(V, E)$ with V as the set of vertices and E as the set of edges, a pattern is the joint feature of every n connected vertices and describes the way how they are linked together. Such a pattern always comprehends all existing edges between n vertices. Furthermore, none of the n vertices is isolated from the others, i.e. each of the n vertices must be directly attached to at least another one.

The total set of distinct n-vertex patterns in G is given by $P^n = \{P_1^n, P_2^n, \cdots, P_i^n\}$. The uniqueness of the i-th pattern is due to its structure and the particular set of n-vertex subgraphs in G, $P_i^n = \{P_{i,1}^n, P_{i,2}^n, \cdots, P_{i,j}^n\}$, whose vertices are exactly connected to each other as described by the pattern. A subgraph $P_{i,j}^n$ is also denoted as the j-th *instance* of pattern P_i^n and there is no other subgraph in G that consists of the same set of vertices and edges than $P_{i,j}^n$ (Figure 1). Importantly, an edge e of a pattern instance $P_{i,j}^n$, $e \in E_{i,j}^n$ where $E_{i,j}^n \subseteq E$, is incident only to vertices of this instance and denoted as an *intrinsic* edge of the pattern $P_{i,j}^n$. Other edges in G, $e \in E \backslash E_{i,j}^n$, do not account for the coherence between the vertices of $P_{i,j}^n$ and are called *extrinsic* edges.

For the global context of a graph G we propose to evaluate the importance of pattern instances based on their participation on the existing connections in G as generally introduced in [3]. More precisely, we estimate how crucial such an instance is for sustaining the connection between the existing pairwise linked vertices in G by destroying the coherence within the instance. The latter is accomplished by removing all *intrinsic* edges of a pattern instance since they essentially reflect a particular pattern. The more pairwise connected vertices in G become disconnected due to the elimination of the *intrinsic* edges the higher is the importance, i.e. topological significance, of a pattern instance $P_{i,j}^n$. This influence is quantified by the *pairwise disconnectivity index* of a pattern instance [4] which is defined as

$$Dis(P_{i,j}^n) = 1 - \frac{N'}{N} \tag{1}$$

Equation 1 depicts the fraction of those initially connected ordered pairs of vertices in G which have become disconnected upon the removal of all intrinsic edges of pattern instance $P_{i,j}^n$. Hence, N is the number of ordered pairs that are linked by at least one path

in G and N' is the number of ordered pairs in $G'(V,E')$ with $E' = E \backslash E^n_{i,j}$. G' is thus the subgraph of G that results from removing the intrinsic edges of the pattern instance $P^n_{i,j}$ from G.

The topological significance of a pattern P^n_i is determined by the topological impact of multiple representatives of this pattern on the connectivity in G. The respective influences of all of its instances need to be considered equally to avoid a misleading result which might occur by choosing a procedure as in [9]. This can be achieved by averaging over all instances of a pattern, i.e.

$$Dis(P^n_i) = \overline{Dis(P^n_{i,j})} = \frac{1}{J}\sum_{j=1}^{J} Dis(P^n_{i,j}) \tag{2}$$

Similar to Eq. 1, the *pairwise disconnectivity index of a pattern* varies between 0 and 1 whereas $Dis(P^n_i) = 0$ means that none of its J instances is crucial for the connection between any pairwise linked vertices. Consequently, $Dis(P^n_i) = 1$ refers to the case where no pair is connected anymore.

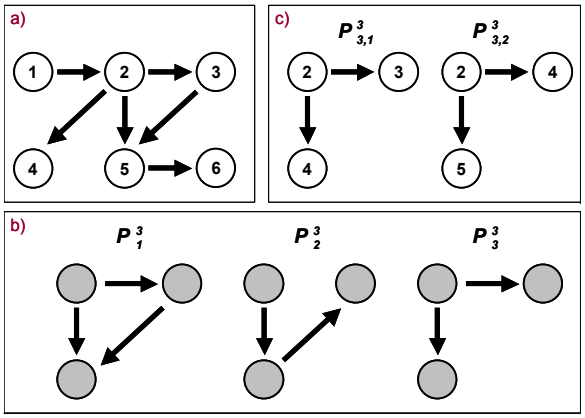


Figure 1: Topological patterns and their instances in a network. (a) A toy network with the set of vertices $V = \{1,\ldots,6\}$; (b) the entirety of all 3-vertex patterns $P_3^3 = \{P_1^3, P_2^3, P_3^3\}$ in the toy network; (c) the two instances of the pattern P_3^3. Each instance is characterized by its individual sets of intrinsic and extrinsic edges. Thus, in regard to instance $P^3_{3,1}$, edges (2,3) and (2,4) are intrinsic and all other edges are extrinsic.

The topological significance of a pattern instance can be comprehended from the example of a feed-forward loop (FFL) included in Figure 1a. The respective FFL-instance is given by the edges $2 \rightarrow 3$, $2 \rightarrow 5$ and $3 \rightarrow 5$. Other edges, that may also start and/or end in the vertices 2,3,5 are the extrinsic edges of this feed-forward loop instance.

Whether this pattern instance may have any influence on the connection between pairwise linked vertices depends on how such a connection is built. If all directed paths between two vertices consist only of extrinsic edges, the pattern instance can hardly have an impact, i.e. the connection is independent from the FFL-instance. For example in

Figure 1a, this is the case for the pair {1, 4}. Note that similar applies even if an intrinsic edge is part of a path that links two vertices but still another path is present that does not contain any intrinsic edge.

Thus, all paths between a connected pair of vertices must contain at least one of the intrinsic edges so that the FFL-instance is critical for this connection. For the example in Figure 1a, this is the case for the pair {1, 6}, which depends on the feed-forward loop instance. The respective pairwise disconnectivity index is 0.67, i.e. the connection of eight of the twelve pairwise connected nodes critically depends on the FFL-instance.

3 Results and Discussion

3.1 Autoregulation as a Feature of the Most Important Nodes

The simplest topological pattern consists of one node and one self-edge, i. e. a gene or molecule is acting on itself: In the transcription network, such self-looping is provided by a TF-gene, the product of which binds to its own promoter; in the signaling network, a signaling molecule may "autocatalytically" activate itself (in most cases, the subunits of a homomeric complex cross-activate each other); in the gene-centric view of the metabolic network, a self-loop usually represents a reversible reaction (an enzyme consumes its own product).

For self-loops, we focus on the properties of the respective nodes since a self-edge cannot have a topological impact on the network coherence. Therefore, we compare the nodes with and without self-loops in the three networks based on their frequencies, inout-degrees, betweenness centrality and the pairwise disconnectivity index applied on vertices. Particularly the latter two metrics are useful for estimating the impact onto a whole network, but depict different properties and maybe complementarily used in topological analyses [3]. In contrast to betweenness centrality that considers the total number of shortest paths going through the given vertex [10][11], the pairwise disconnectivity index does not imply any simplifying assumptions about the significance of paths' length, but rather quantifies how crucial the given vertex is for sustaining the communication ability between connected pairs of other vertices in a network [3].

Table 1. Comparison of nodes with/without self-loops in different mammalian networks.

	Transcription		*Signaling*		*Metabolic*	
	Self-loop	*Other*	*Self-loop*	*Other*	*Self-loop*	*Other*
Frequency	63	216	53	1518	94	1699
Z-Score	39.4	-	34.3	-	51.6	-
Mean inout-degree	10.2	3.1	16	3.9	17.5	5.5
Mean betweenness	0.0021	0.0002	0.004	0.0006	0.0025	0.0005
Mean disconnectivity	0.0261	0.0085	0.0056	0.0021	0.0043	0.0017

The total number of nodes with self-loops in these mammalian networks (Table 1) significantly exceeds the number that might be expected in the corresponding random network of the same size. The high values of Z-score calculated as it was suggested in [12] indicate that self-loops are a network motif there.

These three networks clearly differentiate: First, the relative frequency of nodes with self-edges is significantly higher in the transcription than in the two other networks (29% vs. 3.5% and 5.5%, respectively). Second, these nodes also show a significantly higher mean inout-degree, betweenness centrality and pairwise disconnectivity index. The betweenness centrality is most distinctive in the transcription network: about 10.5-fold more shortest paths pass each autoregulatory node (on average) than each node that has no self-edge in the transcription network, whereas the ratios are 6.7- and 5-fold for signaling and metabolic networks, respectively. The corresponding ratios for the pairwise disconnectivity index are 3.1-fold in transcription, 2.7-fold in signaling and 2.5-fold in metabolic networks. The picture becomes even clearer when considering the maximal values observed: the self-regulating node with maximal betweenness centrality or pairwise disconnectivity index exhibits a much higher value than the corresponding node without self-edge in the transcription network. In contrast, these maximum values are the same, or even lower, for autoregulatory nodes in the other two networks. Altogether, autoregulation seems to be a property of a node that directly goes along with a high topological importance of this node in a network.

3.2 The Mutual Regulation of Two Nodes is a Motif

Two kinds of very basic regulation are conceivable in two-node patterns: first, one node is under the control of another one, represented by a directed edge; second, mutual regulation of the two nodes as indicated by two edges in opposite direction ("binary loop"). Surprisingly, the second pattern is found more frequently than it would be expected by chance, as indicated by the positive *Z*-scores, and therefore is a motif in all three networks (Table 2). The first pattern is correspondingly "under-represented".

Table 2: Two-node patterns in different mammalian networks.

	Transcription			*Signaling*			*Metabolic*		
Pattern	*Freq.*	*Z-Score*	$\overline{Dis}$	*Freq.*	*Z-Score*	$\overline{Dis}$	*Freq.*	*Z-Score*	$\overline{Dis}$
$\bullet \longrightarrow \bullet$	576	-7.73	0.0026	3316	-15.88	0.0004	5212	-73.5	0.0002
$\bullet \longleftrightarrow \bullet$	18	**7.73**	0.0032	55	**15.88**	0.0015	232	**73.5**	0.0007

The column *Pattern* depicts the respective pattern, the column *Frequency* gives the number of occurrences of a pattern in a network, a positive *Z-score* indicates statistically significant over-representation. The column $\overline{Dis}$ gives the mean *pairwise disconnectivity index* of all instances of a pattern.

The highest over-representation is found in the metabolic network, which indicates the presence of many antagonistic enzymatic activities such as kinase-phosphatase pairs, where one enzyme consumes (e.g., a phosphoric acid ester) what the other produces (e.g., the free alcoholic hydroxyl group) and *vice versa*. The binary loop between two neighboring nodes in signaling pathways is much rarer than in metabolic pathways, and may indicate mere undirected physical interactions, e.g. between heteromeric subunits of a receptor, or cross-activations between such subunits, etc. The feed-back activation (or repression) in the transcriptional network, also occurs more frequently than statistically expected, but to a lesser extent than in the other two networks.

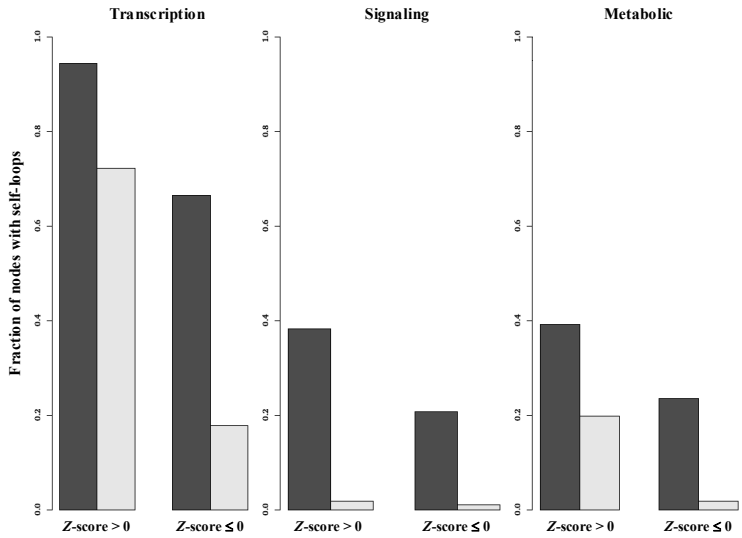


Figure 2: The occurrence of self-loop nodes among 2-node patterns in different mammalian networks. For each network, the patterns (shown in Table 2) are divided according to their frequency: *Z*-score > 0 (over-representation) and *Z*-score ≤ 0. Black and gray bars give the percentage of instances that include at least one node with a self-loop or exactly two self-loop nodes, respectively.

The mean *pairwise disconnectivity index* values are highest in the transcription network, probably due to its smaller size. It is noticeable, however, that these values differ much more between the two 2-node patterns in the metabolic (with a ratio of the $\overline{Dis}$ value for the one-sided pattern to the binary loop of 3.5) and signaling network (ratio 3.8) than in the transcription network (ratio 1.2).

When analyzing the occurrence of self-loop nodes among these two 2-node patterns, we have observed that appear much more frequently in the motif (the 2-node loop, see Table 2) than in the non-motif pattern (or even "anti-motif"), the one-sided pattern (Figure 2). This effect is even more pronounced when comparing motifs and anti-motifs where both nodes possess a self-loop. The higher frequency of self-loops in the transcription network corresponds to a higher percentage of 2-node patterns with self-loops in the transcription network compared with the other two networks. However, the enrichment of self-loops attached to both nodes of binary loops in the transcription

and the metabolic network is evident, whereas it is much lower in the signaling network (Figure 2).

For the metabolic network, this observation may indicate that there are many mutually connected reversible enzymatic reactions so that whole parts of metabolic paths are reversible. This coincides with our knowledge about biochemical pathways which frequently comprise only few reactions that are *de facto* irreversible under the thermodynamic constraints of a living cell. Particularly interesting seems the observation for the transcription network, where the mutual control of two TFs is frequently accompanied by autoregulation of one or both of the participating TFs. This finding deserves further investigation.

3.3 Three-Node Patterns in the Networks Analyzed

We next analyzed the three networks for the recently most intensively studied kind of patterns, i.e. patterns made of three vertices. Of particular interest here is the feed-forward loop (FFL) pattern which has been shown previously to be a motif in several transcription networks [4]. By computing the frequency of FFL patterns in the networks analyzed here, we obtained significantly positive *Z*-scores for this pattern in the transcription, signaling and metabolic network (FFL, pattern 38 in Table 3). The same is true for derived patterns 46 and 166, each of them showing one mutual interaction between two of the three nodes and may thus be comprehended as two superimposed FFLs. Also pattern 102 may be considered as motif in all three networks, although its *Z*-score is very low in the transcription network; this pattern may be viewed as one FFL superimposed with one feed-back loop.

In contrast, the feed-back loop (FBL, pattern 140) is a motif only in the signaling and the metabolic network, but not in the transcription network. The low frequency of FBL in the *E. coli* transcription network was already reported by Alon and colleagues [1], and was observed by us additionally for the transcription networks of yeast and mammals [4]. This may also be a reason why the superposed FFL-FBL motif (pattern 102) has such a low *Z*-score in the transcription network.

Other observations may be more likely explicable by lack of present knowledge rather than reflect genuine features of the respective network. For instance, the high *Z*-score of pattern 36 (two nodes acting on a third one) in the metabolic and the low *Z*-score of the same pattern in the signaling network maybe real: There are certainly many metabolic enzymes that accept different substrates, produced by different enzymes, or one and the same substrate is produced by several other enzymes. In contrast, in signal transduction, convergent information flows resulting in a similar pattern topology are statistically under-represented, although there is a considerable number of such instances (13,606) in the network analyzed. However, we have to assume that the transcriptional regulation of TF-genes, like that of any other gene, is normally exerted by more than just one (other) transcription factor. Therefore, the statistical under-representation of pattern 36 in the transcription network is most likely due to the lack of knowledge about many of

the edges that exist in reality.

It is of interest that a common feature of all those patterns that we have identified as motifs (Z-score > 0) in all three networks, is that all their nodes are connected with each other thus forming loop structures of different configurations.

In general, the mean pairwise disconnectivity index indicates that on average only a few percent (mostly around or below 1%; up to 2.2% in case of pattern 102) of all pairwise connections are disrupted when taking out one of these pattern instances (Table 3, $\overline{Dis}$). The average impact of any 3-node pattern is higher in the transcription than any of the two other networks, which may be mainly due to its smaller size. This shows that the networks are rather robust. In most cases, by removing the intrinsic edges of an individual pattern instance, only a very limited number of the existing connections in a network are destroyed and no strong impact at the global scale is observed.

Table 3: Three-node patterns in different mammalian networks.

		Transcription			*Signaling*			*Metabolic*		
Pattern	*ID*	*Freq.*	*Z-Score*	$\overline{Dis}$	*Freq.*	*Z-Score*	$\overline{Dis}$	*Freq.*	*Z-Score*	$\overline{Dis}$
	6	1916	-0.39	0.0023	11774	-8.92	0.0007	10390	-82.38	0.0003
	12	1068	-1.67	0.011	11865	-9.67	0.0009	14208	-44.75	0.0009
	14	73	-10.47	0.0079	881	-9.48	0.0024	1118	-34.37	0.0016
	36	1620	-2.91	0.0044	13606	-6.91	0.0005	47485	102.03	0.0002
	38	129	5.68	0.0050	496	14.24	0.0006	1627	68.15	0.0003
	46	17	11.18	0.0042	29	8.76	0.0001	85	27.47	0.0003
	78	4	-10.85	0.0099	49	-5.41	0.0011	336	-69.02	0.0014
	102	3	0.35	0.0224	23	8.78	0.0018	101	29.04	0.0044
	140	1	-0.88	0.0137	38	4.64	0.0009	29	6.27	0.0029
	164	197	-6.52	0.0051	722	-10.15	0.0012	4228	-69.65	0.0009
	166	20	7.12	0.0058	21	8.38	0.0001	492	82.09	0.0003
	174	6	7.31	0.0109	9	6.83	0.0001	71	23.75	0.0005
	238	1	-	0.0073	-	-	-	49	729.30	0.0001

The column *Pattern* depicts the respective pattern, to each of them an identifier (*ID*) was assigned. The column *Frequency* gives the number of occurrences of a pattern in a network, a positive *Z-score* indicates statistically significant over-representation. The column $\overline{Dis}$ gives the mean *pairwise disconnectivity index* of all instances of a pattern. The values for the transcription network have been taken from [4].

For none of the networks, a positive correlation between the mean *pairwise disconnectivity index* and the *Z*-score of a pattern could be observed, i.e. motifs are not *per se* more important for the overall network coherence than patterns that are not over-represented.

In spite of this unremarkable feature of the average values, individual instances can be identified for each pattern that show a remarkably high *pairwise disconnectivity index*. In this respect, non-motif patterns exhibit at least as many outliers as motifs (Figure 3). This has already been observed earlier for transcription networks [4], and has been confirmed here for signaling and metabolic networks (Figure 3). It is worth to note that nearly all 3-vertex pattern instances with the highest topological significance in the metabolic network are non-motif patterns and do not exhibit any of the loop structures.

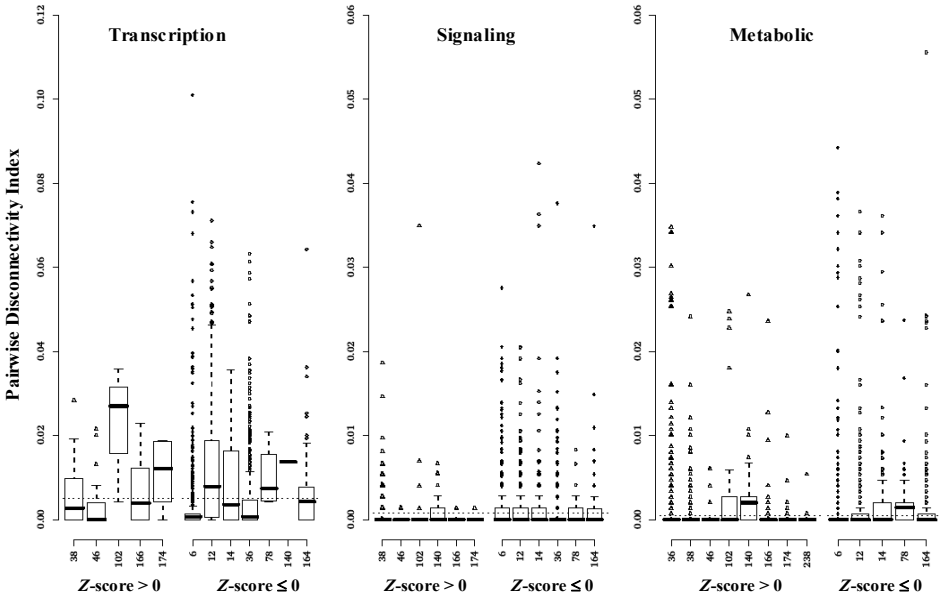


Figure 3: Topological significance of 3-vertex pattern instances in various mammalian networks. For each network, the left and right group of boxplots give the distribution of the *pairwise disconnectivity index* among the 3-vertex motifs or patterns that are not over-represented, resp. The patterns are denoted by their *ID*s (Table 3). The dotted line indicates the average *pairwise disconnectivity index* of all pattern instances in the corresponding network. Note that one point of outliers may represent several pattern instances.

As already described for the 2-node patterns (see 3.2), we also observed an enrichment of vertices with self-loops in 3-node patterns, in particular for the transcription network, but also for the metabolic networks (Figure 4). The difference to the signaling network becomes particularly obvious when focusing on those 3-node patterns where two or even all three vertices have a self-loop. Consistent with our observations about a preferential inclusion of self-loops with binary loops (Figure 2), all 3-node patterns that comprise such a binary loop seem to be particularly rich in self-loops as well.

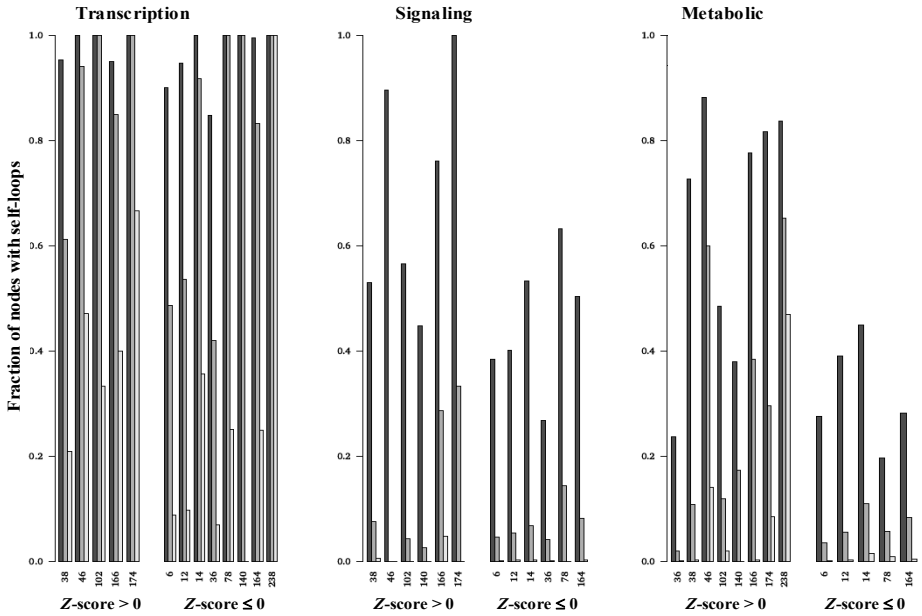


Figure 4: The occurrence of self-loop nodes among 3-node patterns in different mammalian networks. For each network, the patterns (denoted by their IDs according to Table 3) are divided according to their frequency: *Z*-score > 0 (over-representation) and *Z*-score ≤ 0. Black, gray and white bars give the fraction of instances that include at least one, at least two or exactly three self-edge vertices, respectively.

3.4 Large Important Subnetworks Derived from Pattern Analysis

We have noticed that many genes / nodes are particularly frequently re-used among those pattern instances that exhibit a very high *pairwise disconnectivity index*. Most of these instances form subgraphs with interesting biological features:

The largest subgraph thus identified in the transcription network comprises 15 TF-genes (Fig. 5a). A number of them are known to be involved in proliferation (*E2F1*, *NSEP1*, c-*myc*, *HMGA1*, c-*fos*, N-*myc*, *POU1F1*) and/or differentiation events (*CEBPA*, *PAX3*, *MITF*, *RUNX2*, *POU1F1*), altogether targeting at the c-*fos* protooncogene. Also, three nuclear steroid receptors are part of this subgraph (*NR3C1* / glucocorticoid receptor, *NR1I3* / constitutive androstane receptor, *ER1* / estrogen receptor 1), rendering it probable that the cell cycle regulatory effects this subnetwork may exert are also under some hormonal control. From c-*fos*, only two edges point back to other TF-genes, namely to c-*myc* and *HMGA1*. These three genes form one of the three instances of pattern 102, the superimposed FFL/FBL motif. It seems noteworthy that 7, i.e. nearly half, of the 15 nodes of this subgraph exhibit autoregulatory edges (Fig. 5a). These are twice as many as in the whole transcription network, where only 29% of all nodes exhibit a self-edge

(Table 1).

In the signaling network, the largest connected subgraph composed of three-node patterns with high disconnectivity is the one depicting the neighborhood of p53, the central regulator of cell cycle and apoptosis (Fig. 5b). Here as well, 11 out of the 12 connected molecules are involved in cell cycle regulation (Cdk1, cyclin B, APC2, p300, Plk1, Pin1, securin, RSK2, p53, Bcl-xL, Aurora-A). Five out of the 19 edges in this subnetwork are known to have a negative sign, i.e. represent an inhibition. Only one node, RSK2, possess an autoregulatory edge (~8%, compared with the 3.5% in the whole signaling network), which might be in the range of what was to be expected. RSK2 is known to autophosphorylate its serine 386 [13].

Concatenation of three-node patterns with high disconnectivity from the metabolic network does not lead to any prominent subnetwork. The only one is a tree-like structure where nucleoside-diphosphate kinase is the "donor" and many NTP-consuming enzymes are targets (not shown).

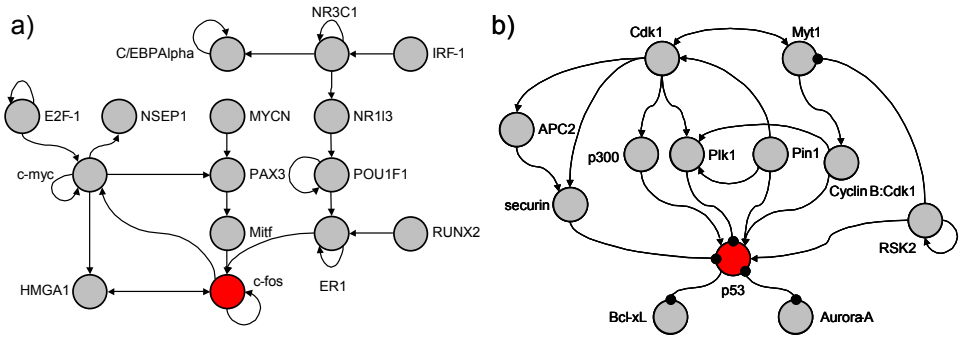


Figure 5: Subnetworks obtained from concatenation of patterns with highest *pairwise disconnectivity index* and common nodes/edges. (a) From the transcription network, a subgraph around the c-*fos* protooncogene was obtained; in addition, known self-loops have been included; (b) from the signaling network, a p53-centered subnetwork was retrieved; line with a dot-end represent inhibitory arcs, the arrows indicate activating interactions.

4 Conclusions

Enrichment analysis and assessment of the topological importance of small network patterns have been proven to complementarily reveal interesting properties of topological patterns and their specific instances. The three kinds of mammalian networks investigated in this study differ characteristically with regard to content and role of their patterns.

The 1-node pattern with a self-edge, termed here "self-loop", has been found to occur in a particularly high frequency in transcription networks. Nodes with a self-edge on average exhibit a higher *inout*-degree, a higher *betweenness centrality*, and a higher *pairwise disconnectivity index* than vertices without a self-loop. All 2- and 3-node patterns that are over-represented and thus qualified as motifs in all three networks exhibit a loop structure. This intriguing observation can be viewed as an advantage of loop-like structures in building up the regulatory circuits of the whole cell. It is in

accordance with the expected role of the loop structure in synchronizing the dynamical processes in scale-free networks [14].

Analysis of the two 2-node patterns relevant in the networks analyzed here also revealed that the binary loop is over-represented, whereas the linear pattern is under-represented. Likewise among the 3-node patterns, all over-represented structures (i.e., motifs) exhibit a loop structure. Both the 2- and the 3-node loops are additionally enriched by attached self-loops, suggesting that loop structures have adopted an important role during evolution of these networks.

The previously introduced *pairwise disconnectivity index* [3] has proven useful to provide a metric of the importance of individual topological patterns for the coherence of the network. While abundance and average importance of patterns may go along with each other in the case of 1- and 2-node patterns, no positive correlation has been observed for 3-node patterns. Non-motif patterns exhibit at least as many outliers with highly elevated *pairwise disconnectivity index* as motifs do. 3-Node pattern instances with the highest pairwise disconnectivity values usually do not belong to any of the loop structures.

Some of these 3-node patterns showing highest importance for the respective network revealed overlapping components and could be concatenated to larger subgraphs. This was possible for the transcription and the signaling network, and both subgraphs turned out to be in connection with the regulation of cell proliferation. We regard this as prove for the biological relevance of our disconnectivity analysis.

Summarizing the comparison of the different networks, we have noticed that the transcription network differs from the other networks in that (*i*) self-loops play an even higher role, (*ii*) its binary loops are highly enriched with self-loops attached, and (*iii*) feed-back loops are not over-represented. Metabolic networks reveal some particular topological properties which may reflect the fact that metabolic paths are, to a large extent, reversible.

Acknowledgments

Part of this work was funded by a grant from the European Community Sixth Framework Program (FP6) under grant agreement number 037590 (MA, Net2Drug project) and by the EurotransBio project GlobCell (BG, grant no. 0315225B).

References

[1] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., Network motifs: Simple building blocks of complex networks, *Science*, 298:824-827, 2002.

[2] Shalgi, R., Lieber, D., Oren, M., and Pilpel, Y., Global and local architecture of the mammalian microRNA-transcription factor regulatory network, *PLoS Comput. Biol.*, 3(7):e131, 2007.

[3] Potapov, A.P., Goemann, B., and Wingender, E., The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks, *BMC Bioinformatics*, 9:227, 2008.

[4] Goemann, B., Wingender, E., and Potapov, A.P., An approach to evaluate the topological significance of motifs and other patterns in regulatory networks, *BMC Syst. Biol.*, 3:53, 2009.

[5] Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., and Wingender, E., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res*., 34(Database issue):D108-D110, 2006.

[6] Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kroneberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O., and Wingender, E., TRANSPATH®: An information resource for storing and visualizing signaling pathways and their pathological aberrations, *Nucleic Acids Res.*, 34(Database issue):D546-D551, 2006.

[7] Choi, C., Crass, T., Kel, A., Kel-Margoulis, O., Krull, M., Pistor, S., Potapov, A., Voss, N., and Wingender, E., Consistent re-modeling of signaling pathways and its implementation in the TRANSPATH database, *Genome Inform.*, 15(2):244-254, 2004.

[8] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y., KEGG for linking genomes to life and the environment, *Nucleic Acids Res.*, 36(Database issue):D480-D484, 2008.

[9] Dobrin, R., Beg, Q.K., Barabási, A.L., and Oltvai, Z.N., Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network, *BMC Bioinformatics*, 5:10, 2004.

[10] Freeman, L.C., A set of measures of centrality based on betweenness, *Sociometry*, 40:35-41, 1977.

[11] Girvan, M. and Newman, M,E., Community structure in social and biological networks**,** *Proc. Natl. Acad. Sci. USA*, 99: 7821-7826, 2002.

[12] Alon, U., *An Introduction to Systems Biology. Design Principles of Biological Circuits*, Chapman & Hall/CRC, Boca Raton, 2007.

[13] Frödin, M., Jensen, C.J., Merienne, K., and Gammeltoft, S., A phosphoserine-regulated docking site in the protein kinase RSK2 that recruits and activates PDK1, *EMBO J.*, 19(12), 2924-2934, 2000.

[14] Ma, X., Huang, L., Lai, Y.C., and Zheng, Z., Emergence of loop structure in scale-free networks and dynamical consequences, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys*., 79(5 Pt 2):056106, 2009.

Anhang B

Artikel 4 aus dem Jahr 2011 in Bioinformatics (zur Veröffentlichung eingereicht)

Application Note

DiVa online: Finding key elements in regulatory networks by means of the pairwise disconnectivity index

B. Goemann*, T. Schoeps, A. Potapov and E. Wingender

Department of Bioinformatics, University Medical Center, Georg August University Goettingen, Goldschmidtstrasse 1, 37077 Goettingen, Germany

ABSTRACT

Summary: We provide *DiVa online*, a web server for the automated analysis of regulatory networks by means of the pairwise disconnectivity index (PDI). The PDI is a metric that ranks the importance of vertices, edges, or groups of them (like patterns) based on their contribution to the connectedness of a network. Our service may be used for a targeted analysis as well as for screening a network for the most interesting candidates. It may also be utilized for the detection and analysis of topological patterns (e.g. motifs) in regulatory systems.

Availability: The web service is free and can be accessed at http://diva.sybig.de. JavaScript must be activated in your web browser.

Contact: bjoern.goemann@bioinf.med.uni-goettingen.de

1 INTRODUCTION

In the last years, graph theoretical approaches were increasingly recognized as a powerful toolbox for the analysis of biological networks. They helped to identify the architectural peculiarities of various networks and greatly advanced the understanding of their underlying functionalities. However, most of these studies concentrate on the large-scale features of biological systems, so that the variety of approaches for detecting the key elements in them is rather limited. In contrast, experiments typically deal with the involvement of individual genes/molecules in the context of interest. Reliable theoretical analysis techniques and appropriate, efficient calculation services are thus very much needed.

2 METHODS

In this regard, we have recently introduced a new metric - the *pairwise disconnectivity index* (PDI) [1]. The PDI rates the global importance of an entity (vertex, edges or groups of vertices/edges) by estimating how essential it is for the connectedness of a network. The way this is accomplished is quite similar to how the role of a gene is evaluated in the lab where a gene is knocked out and the arising effect is pursued. The PDI simulates the elimination of an entity and measures whether this completely disrupts the existing connections between any two nodes in the respective network, e.g. gene *A* is no longer linked to gene *B* by means of any path. The more such connected ordered pairs of nodes become disconnected upon the elimination of an entity the more important the entity is for the connectivity of the network.

The PDI generally benefits from that it can be applied on a network without making any preceding simplifications or assumptions about the applied semantics, i.e. the network can be used as it is. The performance of the approach has been demonstrated at first for detecting key nodes in various networks from different kingdoms [1]. The PDI successfully identified well-known key regulator genes which are associated with a function that was proven experimentally to be vital for the respective organism. With this regard, it proved to be superior to the popular betweenness centrality method [1].The concept of the PDI has been further adopted for analyzing network patterns like motifs and was applied to patterns that consist up to three vertices in several networks from different species [2,3].

3 RESULTS AND CONCLUSIONS

Here, we present a calculation service, *DiVa online,* that provides an easy-to-use interface for analyzing networks with the aid of the PDI. In a simple step-by-step procedure (Figure 1) the user may upload a network as adjacency list in text file format and choose one of the following options: (i) scan the network for the most interesting candidates, (ii) calculate the PDI of each vertex/edge, or, (iii) detect and evaluate the significance of network patterns. Subsequent knockout analysis is carried out in the background by the *DiVa* console program, which can also be downloaded for separate use [4]. Pattern and motif detection is done by the FANMOD software in the background [5].

When *DiVa online* has finished the analysis, results are available as a zipped file archive in the download area. Optionally, the user receives a notification by mail with a link to the download area. The archive file contains a readme file that describes the chosen processing method and lists the further contents of the archive. If the user has chosen to compute the PDI for vertices or edges, these include the original network file, and a tab delimited text file with the calculation results, ranking the nodes or edges according to their calculated importance for the whole network. If the choice was to use the PDI for patterns, the original network file, the standard and optional dump output of the FANMOD software (two text files), two tab delimited text files with the calculation results and two PDF files with preliminary statistics are provided.

Detailed instructions on how to use the web service as well as three well-known biological networks are available for an exemplary usage: the networks of (i) transcriptional regulation in *E. coli*, (ii) transcriptional regulation in *S. cerevisiae* and (iii) neuronal connections in *C. elegans*. Key node analysis with the PDI has been published for all three networks [1], exhaustive pattern detection and analysis was reported for the two transcriptional networks [2].

*To whom correspondence should be addressed.

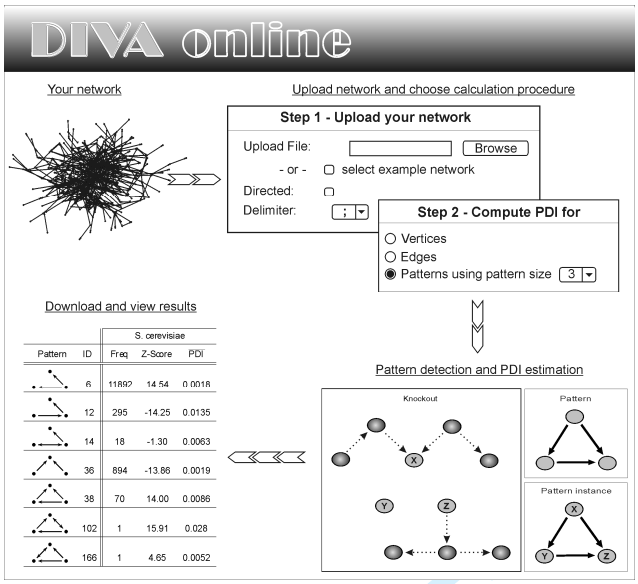


Figure 1: Workflow of *DiVa online* for determining the topological significance of network patterns with the pairwise disconnectivity index

ACKNOWLEDGEMENTS

Funding: This work has been supported in parts by grant 031U110A (Intergenomics) of the German Federal Ministry of Education and Research (BMBF) and by grant 503568 (COMBIO) within the 6th Framework Programme for Research, Technological Development and Demonstration of the European Commission.

Conflict of interest: None declared.

REFERENCES

1. Potapov, A.P., Goemann, B. and Wingender, E.: The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks. *BMC Bioinformatics* 2008, 9:227
2. Goemann, B., Wingender, E. and Potapov, A. P.: An approach to evaluate the topological significance of motifs and other patterns in regulatory networks. *BMC Systems Biology* 2009, 3:53
3. Goemann, B., Potapov, A.P., Ante, M. and Wingender, E.: Comparative analysis of topological patterns in different mammalian networks. *Genome Informatics Series* 2009, 23: 32-45.
4. http://www.bioinf.med.uni-goettingen.de/services/diva
5. Wernicke, S. and Rasche, F.: FANMOD: a tool for fast network motif detection. *Bioinformatics* 2006, 22: 1152-1153.

Anhang B

Artikel 5 aus dem Jahr 2011 in BMC Systems Biology (zur Veröffentlichung eingereicht)

CompAnalysis_R25

The large-scale organization of mammalian networks with different functionalities: transcription, signal transduction and metabolic networks

Björn Goemann, Edgar Wingender and Anatolij P. Potapov*

Department of Bioinformatics, Medical School, Georg August University of Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany

*Corresponding author

Email addresses:

BG: bjoern.goemann@bioinf.med.uni-goettingen.de

EW: edgar.wingender@bioinf.med.uni-goettingen.de

APP: anatolij.potapov@bioinf.med.uni-goettingen.de

Abstract

Background

Various graph theoretical approaches have been applied so far to characterize the architectural peculiarities of well-defined biological networks, mostly from prokaryotes and unicellular eukaryotes. Still, little is known about the topology of regulatory networks in higher eukaryotes. In this study, we have comparatively investigated three different mammalian networks – on transcription, signal transduction and metabolic processes - with respect to their common and individual topological traits.

Results

The networks have been constructed based on genome-wide data collected from human, mouse and rat. None of these three networks exhibits a pure power-law degree distribution and, therefore, could be considered scale-free. Rather, the degree distributions of all three networks were best fitted by a mixed model of a power law with an exponential tail. As revealed by the dependence of the clustering coefficient $C(k)$ on degree k, the transcription and the metabolic network show, to different extent, elements of a hierarchical modular organization; in contrast, the signaling network does not show such a dependency. The connectivity within each network is rather robust, as is seen when removing individual nodes and computing their pairwise disconnectivity index (PDI) values. Just a few vertices turned out to have a strong impact on the networks' coherence. These nodes are characterized by a broad range of degrees and include some of high-degree (hubs), mid-range and even low-degree vertices. In turn, not all hubs exhibit a high topological significance, showing that the degree alone is not the decisive criteria of a node's importance.

Conclusions

The topological properties of the investigated networks reveal distinct architectures. The transcriptional network exhibits a hierarchical modularity, whereas the signaling network is mainly comprised of semi-autonomous modules. The metabolic network, in contrast, seems to be constituted of a more complex mixture of substructures. In any case, high-PDI nodes are considered to play the major role in interlinking the different modules. We conclude that the subsets of genes and relationships that constitute these networks have co-evolved very differently and through multiple mechanisms.

Background

In the last decade, intensive studies of the global architecture of various real world networks led to the conclusion that most of them share the small-world property in conjunction with a power-law degree distribution [1-5]. A small-world network is characterized by a small average shortest path length between any two vertices and a large mean clustering coefficient when compared with random networks of the same size [6]. A power-law degree distribution is generally perceived as synonym for its scale-freeness. It implies that most vertices have a very small number of links while few others, so-called "hubs", are highly linked, although other degree distributions (e.g. exponential decays) exhibit this property as well. It is an important characteristic of such a topology that it assures an amazing robustness against random failures and is sensitive only to targeted attacks on the hub nodes [1,7,8]. They are thought to connect various modules inside a scale-free network, i.e. sets of highly interlinked vertices, which themselves have been reported to be organized in a hierarchical way [1,9,13]. Most notable, however, is that an evolutionary model has been reported that explains well how networks with a scale-free degree distribution have been generated, i.e. through a preferential attachment of newly emerging vertices to hubs [1,4].

So far, such features have been reported for different types of biological networks in various organisms: they range from metabolic networks in bacteria, archaea and eukaryotes [9-12] over protein-protein interaction networks in yeast and fly [13-16] to some signaling [17,18] and eukaryotic gene expression networks [19-23]. For regulatory networks in higher eukaryotes, only scarce information is available so far about their large-scale characteristics [24-27]. The current knowledge about the global organization of biological networks is therefore almost completely based on the analysis of prokaryotes and unicellular eukaryotes. It remains to be seen whether the

architecture of the corresponding networks in multicellular systems considerably deviates from that in unicellular ones. Since the processes in higher eukaryotes require a much higher regulatory overhead to coordinate the differential gene expression programs in various cell types and tissues one may expect several peculiarities in the corresponding networks.

This work extends our previous studies on selected mammalian networks [25,27] by describing and comparing the general topologies of mammalian regulatory systems. Each of the inspected networks displays a specific functional aspect of a mammalian cell: (*i*) the transcription network, representing the relationships between transcription factor genes, (*ii*) the signaling network, comprising all known signal transduction molecules, and (*iii*) the metabolic network composed of genes encoding metabolic enzymes. Our previous studies have already indicated a peculiarity of the mammalian transcription network compared with the signaling and the metabolic network in being void of 3-node feed-back loops [28]. The focus here is on their degree distributions, modular organization, and robustness against random and targeted perturbations. We show that despite being encoded by the same genomes, these networks significantly differ from one another in their general architectural design.

Results

Genome-wide mammalian transcription, signaling and metabolic networks

The view of mutually affecting biological processes in a network perspective has greatly facilitated our understanding of the underlying functionality and occasionally complex interrelationships. Such networks may represent specific aspects of regulatory functions in a cell in nearly any detail or just outline the very general processes at a rather discrete level. Here, we consider various mammalian networks at the level of their orthologous abstraction [29-31], i.e.

pooling the available information for different mammalian species (mostly human, mouse and rat) and thereby neglecting conceivable particularities between them. We have applied this technique to the genome-wide scale, hence focusing on the regulatory potential within the whole genome and not that of its particular parts expressed in various cell types.

We have studied mammalian networks of three different kinds, each of which is represented as a directed graph without double edges. The network of mammalian *transcription* factor genes was retrieved from the TRANSFAC® database, release 11.3 [29], and the TRANSPATH® database, release 8.3 [30]. In this network, the nodes represent genes that are coding for transcription factors (TF-genes), other kinds of genes were not considered. The edges respond to the genetic interactions between TF-genes, i.e. comprise the expression of each TF-gene and the effects of its product, i.e. trans-activation/-repression of the target TF-genes, etc. (Figure 1). This network consists of 279 nodes and 657 edges.

The TRANSPATH® database [30], release 8.3, was also used to reconstruct the mammalian signal transduction network. It shows how various signals are relayed from different receptors to target molecules such as transcription factors and metabolic enzymes. For this, we extracted "semantic" reactions only which focus on the essential components between which information is actively forwarded, and did so at the level of "orthogroups" [31]. This network consists of 1571 nodes and 3425 edges.

The mammalian network of genes encoding enzymes of metabolic reactions was reconstructed from the Ligand section of the KEGG database [32] by retrieving all genes encoding metabolic enzyme activity in mammalian (more precisely: human, mouse and rat) systems. Consistently

following a genome-centric view, the nodes represent genes coding for metabolic enzymes, and the edge semantics is to forward a metabolite produced by one enzyme to one that consumes it (Figure 1). This network emphasizes the role of genetically encoded information in metabolic processes and can be viewed as the line graph of the original metabolic graph. We will refer to this one as to *metabolic* network. Its weakly connected part consists of 1793 nodes and 5538 edges.

The small-world property and presence of modules

The general topological features of these mammalian networks are compiled in Table 1. In the networks the average shortest paths length are with 3.4 (transcription network), 6.2 (signaling network) and 6.4 (metabolic network) fairly small as compared to the respective network size. The mean clustering coefficients of 0.07 in the transcription network, 0.02 in the signaling network and 0.09 in the metabolic network are much larger than the expected values of randomized networks of the same size (Table 1). An increased clustering is indicative for a modular style of network architecture [1,9,]. Following the criteria of small-world networks stated as increased clustering features and small mean shortest path lengths [6], all three mammalian networks are of small-world.

Degree distribution does not show scale-freeness

Recent attempts in reassessing the degree distributions of various networks indicate that in many cases, data were fitted prematurely to a power law (PL), $P(k) \propto k^{-\gamma}$, [33,34]. The reason for wrongly assigning a power-law can be understood from the way how the degree distribution is usually analyzed [34,35]. In particular, the most frequently applied method - linear regression on a double logarithmic plot - suffers from several major drawbacks and may lead to misleading

results [34]. In fact, by using other statistical approaches such as the maximum likelihood method, many of the apparently power-law networks were shown to be of single or broad scale [33,36-38]. Single-scale networks are characterized by a connectivity distribution that decays according to an exponential law (EL), $P(k) \propto e^{-\lambda k}$. In contrast, broad-scale networks display a mixed connectivity distribution that exhibits a power law with an exponential tail (PLET), $P(k) \propto k^{-\alpha} e^{-\beta k}$ [33]. Because the exponential decay term $e^{-\beta k}$ overwhelms the power decay term $k^{-\alpha}$ at large *k*, this distribution is not asymptotic to a power law.

We have analyzed the distributions of the incoming (*k-in*), outgoing (*k-out*) and total (*k-inout*) degrees of the mammalian networks with respect to these three types of models. The fitting parameters were obtained by using a standard maximum likelihood estimators' procedure (see Methods section). Figure 2 depicts the degree sequences for the incoming, outgoing and total degrees of the networks altogether with the fitted power law, exponential law and the power law with exponential tail. The maximum likelihood estimators for the respective parameters of the models γ (PL), λ (EL), α and β (PLET) are summarized in Table 2. As can be seen from the plots in Figure 2, pure PL and EL models fit to the observed distributions to different extent, whereas the PLET model always fits well. Only in some cases (all degree distributions of the transcription network and the out-degree distribution of the metabolic network), the pure EL model seems to fit as well as the PLET.

Due to the clear pitfalls of mere visual inspection, we applied a likelihood ratio test which compares the quality of fitness of the three considered models. Table 3 summarizes the differences of the logarithms of the maximum likelihood functions for the three models using their estimated parameters. A positive value indicates that the first of the two models has a higher

outcome of its maximum likelihood function and therefore is the preferred choice compared to the other one, while a negative value denotes the preference of the second model among the two models tested. As assessed above from a visual inspection of the plots in Figure 2, the likelihood ratio test also reveals a PLET model being superior over the two considered alternatives in fitting the degree distribution of each mammalian network examined (Table 3, the second and third columns). For the transcription network, the preference of a PLET over an EL model is minimal, thereby confirming that the degree distributions of this network can be approximated by the latter. A similar notion can be applied to the outgoing degree distribution in the metabolic network. When just comparing PL and EL models, the power-law model is given preference for the signaling network only, but is nevertheless clearly rejected against the PLET model as a plausible fit for its degree distributions. Hence, none of the networks is scale-free.

Clustering coefficient indicates different architectures of the three mammalian networks

Modular organization is a hallmark of biological systems with each module performing its special functional task [39,40]. Modules are sub-networks of different sizes with an enhanced density of internal links and are thought to be arranged in either a semi-autonomous or hierarchical manner [**Error! Reference source not found.**,9]. The semi-autonomous type of organization might be provided by relatively independent groups of interconnected vertices which altogether do not exhibit a dependence of the average clustering coefficient for nodes with degree k, $C(k)$, on the degree k [**Error! Reference source not found.**,9]. In contrast, a hierarchical modular organization implies that small groups of vertices assemble hierarchically into increasingly larger groups with communication between the different highly clustered neighborhoods being maintained by a few hubs [**Error! Reference source not found.**,9,41]. Such nested organization of modules is expected to provide a dependence of $C(k)$ on the degree k in a power-law fashion,

$C(k) \propto k^{-\omega}$, in a double logarithmic plot [1,9,41]. As was shown with artificially designed and highly regular networks constructed by a repeated duplication and integration process of clustered nodes [1,41], a slope of -1 indicates a modular architecture in hierarchical style throughout a network. A power-law dependence of the clustering coefficient on the degree ($C(k) \propto k^{-1}$) has been suggested to be the signature of hierarchical networks [**Error! Reference source not found.**,9,41].

To assess whether the mammalian networks exhibit hierarchical modularity, we checked the dependency of the average clustering coefficient on the degree (Figure 4). In the case of the transcription network, $C(k)$ clearly depends on k. This dependency can be linearly fitted in a log-log plot and, thus, follows a power-law. However, the overall slope of -0.47 is significantly smaller than the expected -1. This might reflect a relaxed hierarchical setup of modules. On the contrary, $C(k)$ is independent from k in the signaling network and their relation thus strikingly fails to express a power-law behavior in double logarithmic scale. Accordingly, hierarchical modularity cannot be detected in the signaling network and the semi-autonomous type of modular organization is likely to dominate instead. In the metabolic network, there is a dependency of $C(k)$ on k in principle, but the scattering of values in the log-log plot (Figure 3) does not allow a plausible approximation by a power-law. Probably, hierarchical and semi-autonomous types of modularity co-exist in this network.

Network robustness and the importance of hub nodes

Next to a modular organization, biological systems fundamentally feature a remarkable robustness [42]. It has been observed that they can generally withstand perturbations, but are sensitive against targeted attacks on the hub nodes [1,7,8,42,43]. In this regard, we have

examined the three mammalian networks to their tolerance against single node knockouts using the methodology introduced in [44] – the pairwise disconnectivity index (PDI). The method has been shown to be a well-suited measure for detecting key nodes in regulatory networks. It evaluates the topological significance of a node depending on its contribution for sustaining the connectivity of a whole network.

Applying the PDI to the mammalian networks we found that random removal of a node has only marginal impact on the connectedness of these networks (Figure 4). The respective mean PDI values are 0.012 (transcription), 0.0023 (signaling) and 0.0019 (metabolic), i.e. deleting a node expectedly disrupts the connection of just 1.2%, 0.23% and 0.19% of the existing paths. However, a small number of nodes are characterized by a PDI value that significantly exceeds the average PDI values (up to 0.21), thereby rendering the networks vulnerable upon a targeted removal (Figure 4).

Interestingly, these key nodes are not necessarily hubs. As Figure 4 shows, hubs associated with a high PDI value surely exist in the mammalian networks, e.g. *c-fos, c-myc* and *p53* in the transcription network. However, there are also hubs the topological significance of which ranges within the same scale as selected low or mid-range degree vertices (e.g. *Sp1* in the transcription network). In particular within the signaling and metabolic networks many examples were found where hubs have just a small or nearly no impact at all (e.g. *Gi* in the signaling and *apyrase* in the metabolic network). On the other side, a strong effect on the connectedness of all three networks was observed for many nodes with much smaller degrees like *IRF-1* in the transcription and *Myt1* in the signaling network. Consequently, and in agreement with [44], no correlation between a

node's importance for the connectedness of a network and its degree can be detected. The degree alone does not seem to be the decisive criteria for the topological importance of a node.

Discussion

Regulation of cellular processes in higher eukaryotes is characterized by an outstanding complexity. Diverse programs, each involving a large number of molecular entities and their interactions, run in different time scales in an apparently independent manner although being closely interrelated in reality. They generate and forward input-dependent outputs to each other, thereby spotlighting their inner machineries, which enable them to react specifically on a certain composition of inputs (signals, hormones, nutrients, etc.). Much effort has therefore been put into understanding the configurations of the basic regulatory mechanisms.

In this context, several architectural features were detected and advocated as generic attributes of biological regulatory networks: Amongst others, these are (*i*) scale-freeness given by a power-law degree distribution [5], (*ii*) hierarchical modularity [41], and (iii) a robust yet fragile design with hubs as the key nodes [7]. Consequently, very different functional systems seem to share fundamental properties, notwithstanding that some of the concepts, namely the prevalence of power-law degree distributions and the importance of motifs [45], are controversially discussed [27, 28,37,38, ,46,47].

Mammalian networks are not scale-free

With regard to these concepts, we have characterized the topologies of three mammalian networks – transcription, signaling and metabolic – to identify both their global traits and the rather specific properties of their individual vertices. None of the degree distributions of these networks could be convincingly approximated by a power-law, which coincides with the findings of recent studies that have (re-)analyzed the degree sequences of several biological networks [33,36-38,47]. Rather, the best fits were obtained by a power-law with an exponential tail (PLET), which associates power-law behaviour for low-degree nodes fading progressively to exponential law as the degree increases [33,34,38]. This signifies that none of the abstracted mammalian networks, as a whole, is scale-free. However, the lack of scale-freeness does not entail the absence of a hub-centric organization of these networks. Despite the notion of network hubs has originally been associated with a power-law degree distribution, it may equally apply to networks with other types of organization, as in the case of the inspected mammalian networks.

For the transcription network, we found that its degree distributions can also be very well approximated by an exponential law. Such a distribution was reported for the *in*-degree of the transcription networks of yeast [37] and bacteria [48,49]. In contrast, the *out*-degree distributions for these networks do not follow an exponential law [37,48,49], which differs from our findings for the mammalian transcription network. These differences may reflect some peculiarities of transcriptional regulation in mammals as compared with that in bacteria and yeast. Besides, the mammalian transcription network studied here did not include any non-TF-genes, while the bacterial and yeast networks [37,48,49] included both TF-genes and non-TF-genes. Conceivably, the core transcription network comprising only TF genes and the extended transcription network, which also includes non-TF target genes, may be differently organized. Preliminary studies of our group support this hypothesis (M. Haubrock *et al*., in preparation).

The signaling network is the only one where in the direct comparison of the two one-parameter models (PL and EL), the power-law (PL) model is superior to the exponential law (EL) model. However, even in this case the hypothesis that this network is scale-free is rejected when comparing with the PLET model. For the metabolic network our data neither supports a power-law connectivity distribution of metabolites [50,51] nor a coinciding *in*- and *out*-degree distribution of metabolites [50]. Its *out*-degree distribution could be well fitted to an exponential law. That is in accordance with the observation that the connectivity distribution of metabolites in separate functional modules is exponential rather than a power law [52].

The mammalian networks differ in their modular organization

The three mammalian networks differ in the dependencies of the mean clustering coefficient $C(k)$ on the degree k. In both the transcription and the metabolic network, $C(k)$ decreases with the degree. With regard to the criteria of hierarchical modularity [1], these two networks, but not the signaling one, are therefore expected to contain hierarchically organized modules. This implies that while the low-degree vertices are part of highly cohesive, densely interlinked clusters, the high-degree vertices are not, as their neighbors have a smaller chance of linking to each other [1,9]. Hence, the high-degree vertices may play the role of bridging the many small communities of clusters into larger and hierarchically organized parts.

In the case of the transcription network, the linear dependency of $C(k)$ on k indicates a largely hierarchical setup of modules in this network. Since the overall slope deviates from the standard -1, this constitution may not be a pure, but rather an "idealized" one, which in reality may be diluted by other components.

In contrast, the mammalian signaling network appears to be predominantly anti-hierarchical: most of its modules relate to one another in a semi-autonomous fashion. This means that signal processing occurs predominantly *within* individual modules and signal propagation *between* different modules through defined interfaces. Since signal transduction is mostly conveyed by protein-protein interactions, this network structure reflects the fact that its edges depend on the presence of one or few usually highly specific binding sites in each of the constituents (nodes). It ensures the necessary specificity of signal transduction from distinct receptors to distinct targets, such as transcription factors, metabolic enzymes etc., and at the same time efficient re-use of some modules for different purposes. To the best of our knowledge, this particular feature of the mammalian signaling network has not yet been reported, although this observation is in agreement with the general functional role of signal transduction systems.

The more complex dependency of $C(k)$ on k in the metabolic network, i.e. the clear lack of a consistent power-law dependence, may suggest that parts of the network are organized in hierarchical modules, while other parts are constituted by semi-autonomous modules. Such a mixed architectural design is likely to refer well to the division of metabolism into anabolism, catabolism and central metabolism which are subdivided further into multiple incoming and outgoing pathways that label the main functional routes (e.g. glycolysis, tricarboxylic acid cycle) and join at different places to form an interconnected network [53]. The co-existence of various types of modular organization may provide an additional advantage: networks with such a fuzzy community structure are expected to be more efficient in executing the represented processes than those with a pronounced community structure [54].

The three mammalian networks differ in their robustness

By performing a knockout analysis of individual nodes, we have shown that the connectivity of the mammalian networks remains almost unaffected when randomly choosing the node. However, targeting one out of a tiny fraction of selected nodes significantly perturbs the associated network. Some of these particularly important nodes are hubs, but some others are mid-range and low-degree vertices. In accordance with our previous results [44], there is no obvious relationship between the topological significance of a node as measured by its PDI, and its degree. Rather, this significance is determined by the role a node plays for the global connectivity of the whole network, while the degree is only a local property.

In particular the transcription network is characterized by a set of nodes that convey high vulnerability. This, together with the very small average shortest path length and the characteristic dependence of the clustering coefficient on the node degree discussed above, characterizes the architectural design of the transcription network as the most compact and unified among the three networks studied here. The network therefore is hierarchically centralized around a limited set of nodes. This peculiarity of the transcription network is reasonable when considering that in higher eukaryotes the transcription of a gene is generally controlled by a combination of several transcription factors that act together and in a cooperative fashion.

Removing a single node exerted only moderate effects on the pairwise connectivity within the signaling and the metabolic network indicating comparable robustness for both these networks. The high robustness of protein-to-protein connections in the signaling network is particularly remarkable because this network is the most sparsely connected one among the three networks

studied here. It means that connections in the signaling network seem to be backed up by many alternative paths, thus reducing the dependency on single nodes. The semi-autonomous modular organization of the signaling network ensures that a local perturbation is restricted to the affected module, leaving the remainder of the network largely operable. These features of the signaling network also respond well to the functional role of this network in relaying regulatory information via signaling pathways. Note that the observed redundancy refers to a genome-wide signaling network. Since various cell types express only particular parts of this network, the actual redundancy of cell-specific networks might be greatly relaxed.

Critical survey

Obviously, the information used to create these networks, as well as our whole current knowledge about these mammalian systems is not complete. We cannot exclude that some of the numeric values in our estimates will not survive while completing the knowledge. In particular, this might relate to the transcription network, as the most incomplete one. But, nevertheless, we believe that the ascertained features of the inspected networks constitute the main trends of the underlying yet uncharted entire systems.

Conclusions

The topological properties of the investigated networks reveal distinct architectures. The transcriptional network exhibits a hierarchical modularity, whereas the signaling network is mainly comprised of semi-autonomous modules. The metabolic network, in contrast, seems to be constituted of a more complex mixture of substructures. In any case, high-PDI nodes are considered to play the major role in interlinking the different modules. We conclude that the

subsets of genes and relationships that constitute these networks have co-evolved very differently and through multiple mechanisms.

Methods

Estimation of the degree distribution

Estimation of the parameters of the power-law, exponential law and power-law with exponential tail distributions have been calculated using the maximum likelihood method. In the case of the power-law, $P(k) \approx k^{-\gamma} / \zeta(\gamma)$ where $\zeta(\gamma)$ is the Riemann zeta function, i.e. $\zeta(\gamma) = \sum_{k=1}^{\infty} k^{-\gamma}$. The likelihood function is given by $L(\gamma \mid k) = \prod_{k=1}^{N} k^{-\gamma} / \zeta(\gamma)$ with N as the maximum observed degree. Estimating γ is then obtained finding zeros of the derivative of the logarithm of $L(\gamma \mid k)$. As for the exponential-law, $P(k) \approx e^{-\lambda k} / C_1$ where $C_1 = \sum_{k=1}^{\infty} e^{-\lambda k}$ is the normalization constant. The likelihood function is $L(\lambda \mid k) = \prod_{k=1}^{N} e^{-\lambda k} / C_1$ and λ can be obtained by maximizing the logarithm of $L(\lambda \mid k)$ in the same way as for the power-law**.** Finally, the probability function for the power-law with exponential tail is $P(k) \approx k^{-\alpha} e^{-\beta k} / C_2$ where in this case $C_2 = \sum_{k=1}^{\infty} k^{-\alpha} e^{-\beta k}$. The respective likelihood function is $L(\alpha, \beta \mid k) = \prod_{k=1}^{N} k^{-\alpha} e^{-\beta k} / C_2$. Both, α and β can be obtained again by finding zeros of the derivative of the logarithm of the likelihood function.

Dependency of the clustering coefficient on the degree

The clustering coefficient of a node v_i measures how many links, n_i , between its neighbors exist in relation to the maximum possible number of such links, i.e. how well its neighborhood is interconnected: $c(v_i) = n_i / k_i(k_i - 1)$. The mean clustering coefficient for vertices with degree *k*, *C*(*k*), is the average clustering coefficient for all nodes with degree *k*. A network has a hierarchical modular organization if *C*(*k*) approximately follows a power-law of kind $C(k) \approx k^{-\omega}$. When $\omega = 1$, modules are arranged in a hierarchical style throughout a network [1]. The power-law approximation of *C*(*k*) can be seen in a log-log plot.

Pairwise disconnectivity index

The pairwise disconnectivity index estimates the importance of a network element for sustaining the connection between pairwise linked nodes [44]. For a vertex *v*, it is given by $Dis(v) = 1 - N'/N$ where *N'* is the number of still pairwise linked vertices when *v* has been removed from the network and *N* is the number of all initially pairwise connected vertices: The pairwise disconnectivity index ranges between *0* and *1* where zero indicates that *v* is not crucial for the connection of any two linked nodes and one means that no two vertices are connected to each other anymore.

Authors' contributions

APP conceived the study. BG carried out the programming and performed the statistical analysis. APP and BG analyzed and interpreted the data and drafted the manuscript. EW coordinated the work and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Michael Ante for providing the metabolic network. BG was supported in part by the GlobCell project funded by the German Federal Ministry of Education and Research (BMBF).

References

1. Barabási A-L, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**:101-113
2. Dorogovtsev SN, Mendes JFF: **Evolution of networks**. *Adv Phys* 2002, **51**:1079-1187
3. Newman MEJ: **The structure and function of complex networks.** *SIAM Review* 2003, **45**:167-256
4. Albert R: **Scale-free networks in cell biology.** *J Cell Science* 2005, **118**:4947-4957
5. Barabási AL: **Scale-free networks: a decade and beyond.** *Science* 2009, **325**:412-413
6. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393**:409-410
7. Albert R, Jeong H, Barabási A-L: **Lethality and centrality in protein networks.** *Nature* 1999, **401**:130-131
8. Albert R, Jeong H, Barabási A-L: **Error and attack tolerance of complex networks.** *Nature* 2000, **406**:378-382
9. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555
10. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc R Soc Lond B Biol Sci* 2001, **268**:1803-1810
11. Ma H, Zeng AP: **Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.** *Bioinformatics* 2003, **19**:270-277
12. Wagner A: **The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes.** *Mol Biol Evol* 2001, **18**:1283-1292

13. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296:**910-913

14. Yook SH, Oltvai ZN, Barabási A-L: **Functional and topological characterization of protein interaction networks.** *Proteomics* 2004, **4**:928-942

15. Jeong H, Mason SP, Barabási A-L, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**:41-42

16. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RLJr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM: **A protein interaction map of *Drosophila melanogaster*.** *Science* 2003, **302**:1727-1736

17. Papin JA, Hunter T, Palsson BO, Subramaniam S: **Reconstruction of cellular signaling networks and analysis of their properties.** *Nat Rev Mol Cell Biol* 2005, 6:99-111

18. Papin JA, Palsson BO: **Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk.** *J Theor Biol* 2004, **227**: 283-297

19. Featherstone DE, Broadie K: **Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network.** *Bioessays* 2002, **24**:267-274

20. Agrawal H: **Extreme self-organization in networks constructed from gene expression data.** *Phys Rev Lett* 2002, **89**:268702

21. Bhan A, Galas DJ, Dewey TG: **A duplication growth model of gene expression networks.** *Bioinformatics* 2002, **18**:1486-1493

22. Carter SL, Brechbuhler CM, Griffin M, Bond AT: **Gene expression network topology provides a framework for molecular characterization of cellular state.** *Bioinformatics* 2004, **20**:2242-2250

23. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes.** *Nature* 2004, **431**:308-312

24. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, Kershenbaum A, Stolovitzky GA, Blitzer RD, Iyengar R: **Formation of regulatory patterns during signal propagation in a mammalian cellular network.** *Science* 2005, **309**:1078-1083

25. Potapov AP, Voss N, Sasse N, Wingender E: **Topology of mammalian transcription networks.** *Genome Inf Ser* 2005, **16**:270-278.

26. Bhardwaj N, Yan KK, Gerstein MB: **Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels.** *Proc. Natl. Acad. Sci. U.S.A.* 2010, **107:**6841–6846

27. Goemann B, Potapov AP, Ante M, Wingender E: **Comparative analysis of topological patterns in different mammalian networks.** *Genome Inf Ser* 2009, **23**:32

28. Goemann B, Wingender E, Potapov AP: **An approach to evaluate the topological significance of motifs and other patterns in regulatory networks**, *BMC Syst Biol* 2009, **3**:53

29. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S., Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel

H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**(Database issue):D108-D110

30. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kroneberg D, Michael H, Schwarzer, K, Potapov A, Choi C, Kel-Margoulis O, Wingender E: **TRANSPATH®: An information resource for storing and visualizing signaling pathways and their pathological aberrations.** *Nucleic Acids Res* 2006, **34**(Database issue):D546-D551
31. Choi C, Crass T, Kel A, Kel-Margoulis O, Krull M, Pistor S, Potapov A, Voss N, Wingender E: **Consistent re-modeling of signaling pathways and its implementation in the TRANSPATH database.** *Genome Inform* 2004, **15**:244-254
32. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh, M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Res*. 2008, **36**(Database issue):D480-D484
33. Amaral LAN, Scala A, Barthélémy M, Stanley HE: **Classes of small-world networks**; *Proc Natl Acad Sci USA* 2000, **97**:11149–11152
34. Clauset A, Shalizi C, Newman MEJ: **Power-law distributions in empirical data**, *SIAM Review* 2009, **51**:661-703
35. Alderson L, Tanaka R, Doyle J, Willinger W: **Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version)**, *Technical Report* 2005
36. Montoya JM, Pimm SL, Solé RV: **Ecological networks and their fragility**, *Nature* 2006, **442**:259-263
37. Guelzim N, Bottani S, Bourgine P, Képès F: **Topological and causal structure of the yeast transcriptional regulatory network**, *Nature Genetics* 2002, **31**:60-63

38. Khanin R, Wit E: **How scale-free are biological networks?**, *Journal of Computational Biology* 2006, **3**:810-818

39. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks.** *Proc Natl Acad Sci USA* 2003, **100**:12123 – 12128

40. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402**:C47-52

41. Ravasz E, Barabási A-L: **Hierarchical organization in complex networks.** *Phys Rev E* 2003, **67**:026112

42. Kitano H: **Biological robustness.** *Nature* 2004, **5**:826

43. Crucitti P, Latora V, Marchiori M, Rapisarda A: **Error and attack tolerance of complex networks.** *Physica A* 2004, **340**:388 – 394

44. Potapov AP, Goemann B, Wingender E: **The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks**, *BMC Bioinformatics* 2008, **9**:221

45. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U.: **Network motifs: Simple building blocks of complex networks.** *Science* 2002, **298**:824 - 827

46. Konagurthu, A., Lesk, A.: **On the origin of distribution patterns of motifs in biological networks.** *BMC Syst Biol* 2008, **2**:73-81

47. Lima-Mendez G, van Helden J:**The powerful law of the power law and other myths in network biology**. *Mol Biosyst*. 2009, 5:1482-1493.

48. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication**. *Nat Genet* 2004, **36**:492–496

49. Balaji S, Babu MM, Aravind L: **Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of *E. coli***. *J Mol Biol*. 2007, **372**:1108–1122

50. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL: **The large-scale organization of metabolic networks.** *Nature* 2000, 407:651-654

51. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci*. 2001, 268:1803-1810

52. Tanaka R: **Scale-rich metabolic networks.** *Phys Rev Lett*. 2005 94:168101

53. Fell DA: **Metabolic networks.** *In: Biological Networks. Complex Systems and Interdisciplinary Science*. Vol. 3 (ed. KϑpΠs F) World Scientific, New Jersey, 163-197, 2007

54. Danon L, Arenas A, Díaz-Guilera A: **Impact of community structure on information transfer**. *Physical Review E* 2008, **77**:036103

Figure legends

Figure 1: The semantics of edges in the mammalian networks studied here. *Transcription*: a link from transcription factor gene i (TFG_i) to transcription factor gene j (TFG_j) consists of the expression of TFG_i and the subsequent interaction of transcription factor i (TF_i) with the promoter of transcription factor gene j (TFG_j). *Signaling*: an edge $P_i \rightarrow P_j$ refers to a causal link from protein i to protein j. *Metabolic*: an edge $EG_i \rightarrow EG_j$ represents a link mediated by a metabolite that is both the product and educt of two adjacent metabolic reactions which are catalyzed by enzymes E_i and E_j that are expressed by genes G_i and G_j, respectively.

Figure 2: The *in*-, *out*- and *inout*-degree distributions in the mammalian networks do not confirm pure scale-freeness. The logarithm of the respective degree is denoted on the x-axis and opposed to the logarithm of its observed probability (y-axis). In each of the plots, the red line depicts the fitted power law to the data, the green curve represents the fitted exponential law and the blue curves stands for the fitted power law with exponential tail. The maximum likelihood estimators for the corresponding parameters are specified in the small boxes.

Figure 3: The dependency of the mean clustering coefficient $C(k)$ on the degree k. (a): Relationships between $C(k)$ and k in non-logarithmic scale. (b): For each of the mammalian networks, the logarithm of the mean clustering coefficient of all vertices with *inout*-degree k is plotted against the logarithm of k. The diagonal line in each figure has slope 1, following $C(k) \sim k^{-1}$, which responds to the case of throughout hierarchical modularity.

Figure 4: The pairwise disconnectivity index of individual vertices versus their *inout*-degrees. Although the topological significance of a vertex (x-axis) positively correlates with the degree (y-axis), this trend is not straightforward. Hub nodes are neither generally crucial for the connectedness of the networks nor are low or mid-range degree nodes constantly playing a minor role in this context. Instead, a strong effect on the connectedness that is far in excess of the average pairwise disconnectivity indices (dotted vertical lines) occurs for nearly all kinds of degrees. The mean pairwise disconnectivity index and the mean *inout*-degree of all vertices in a network are indicated by the horizontal and vertical dotted lines, respectively.

Figure 1

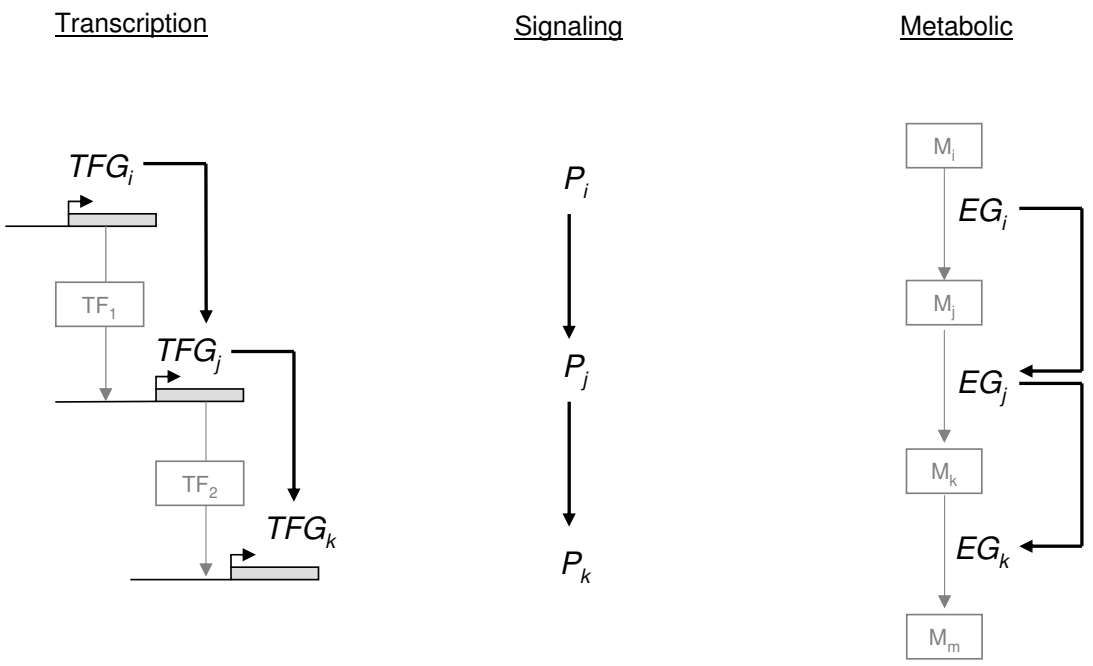

Transcription
Signaling
Metabolic
TFG_i
TF_1
TFG_j
TF_2
TFG_k
P_i
P_j
P_k
M_i
EG_i
M_j
EG_j
M_k
EG_k
M_m

Figure 2

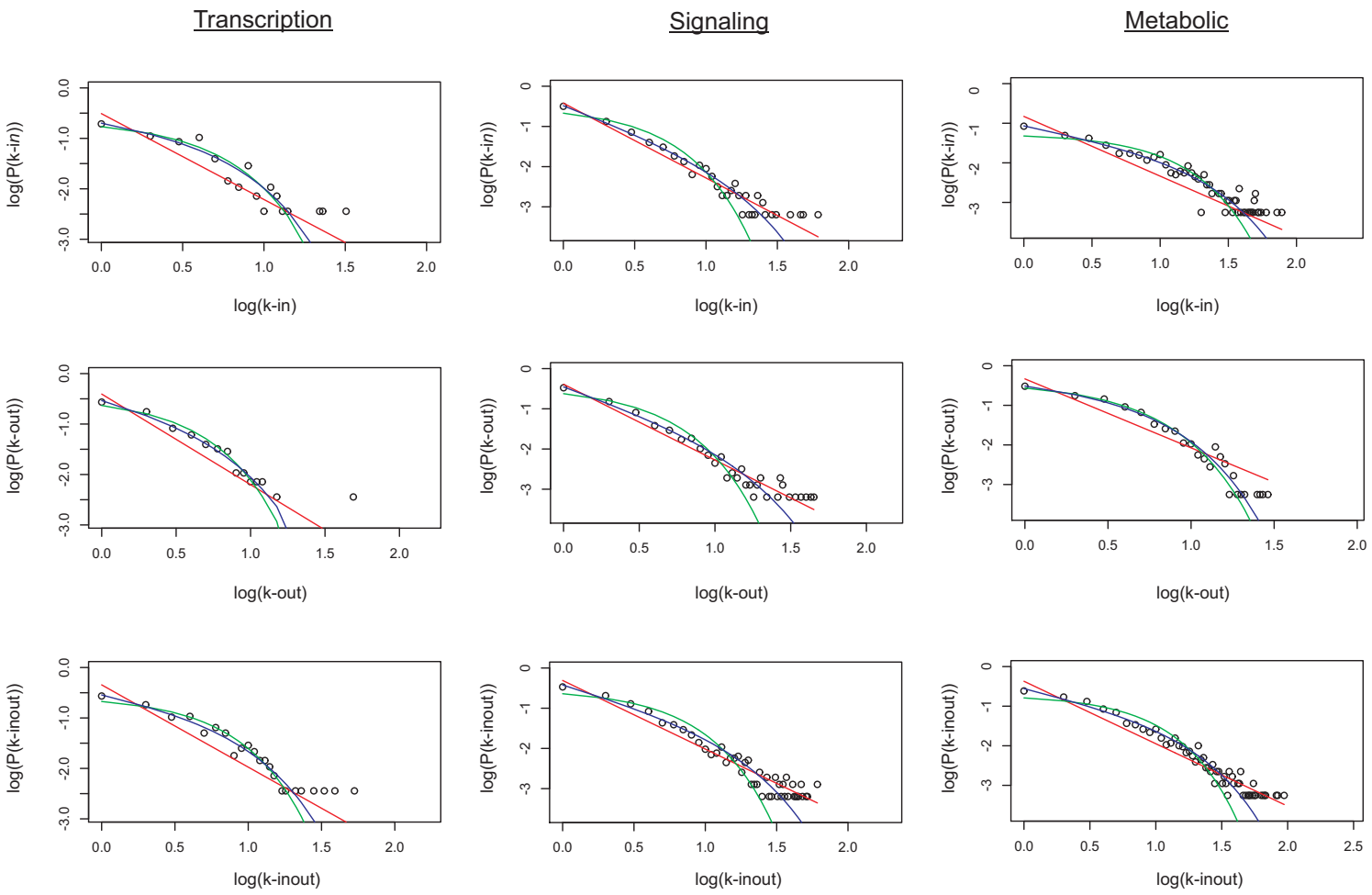

Transcription
Signaling
Metabolic
log(P(k-in))
log(k-in)
log(P(k-out))
log(k-out)
log(P(k-inout))
log(k-inout)

Figure 3

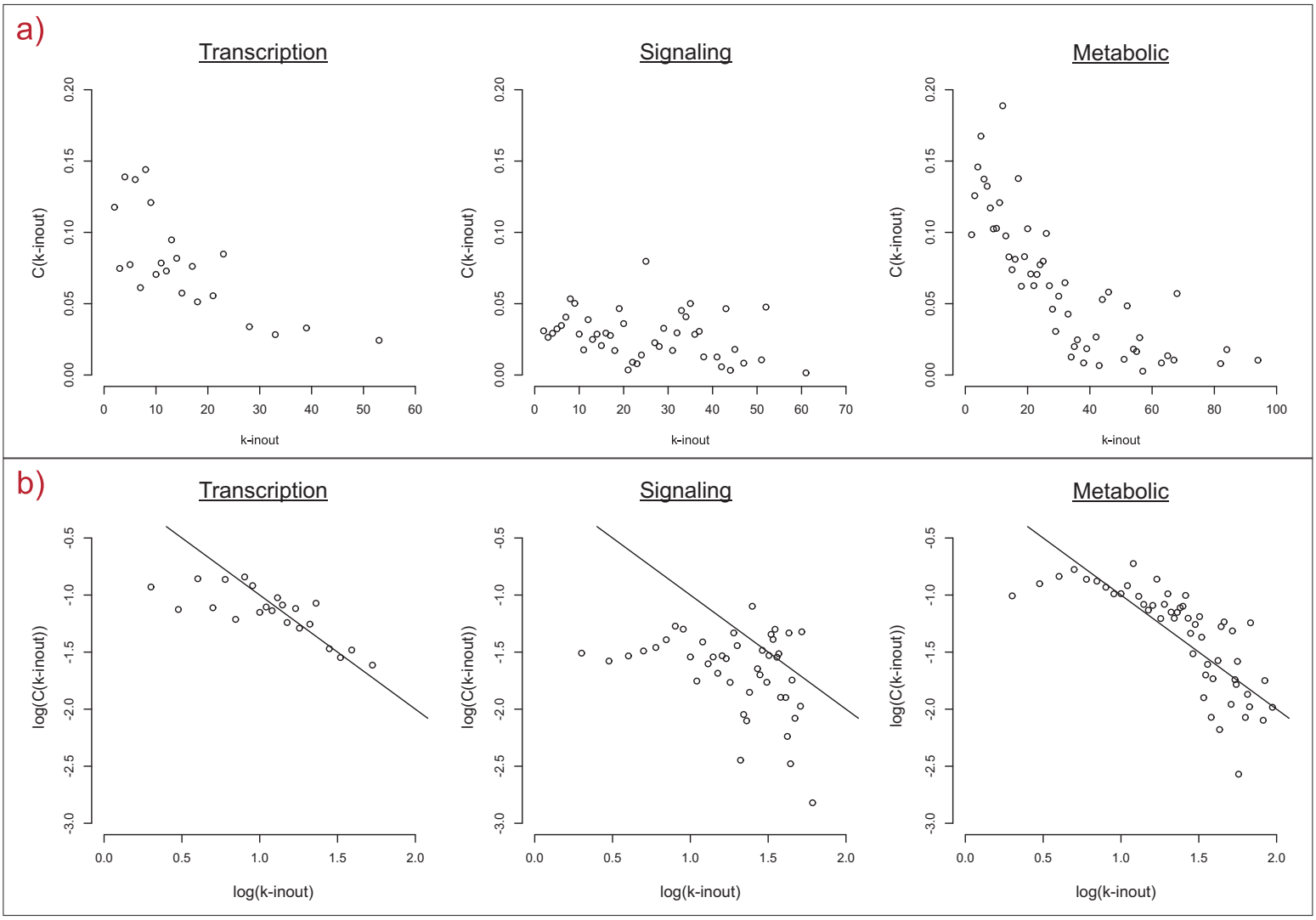
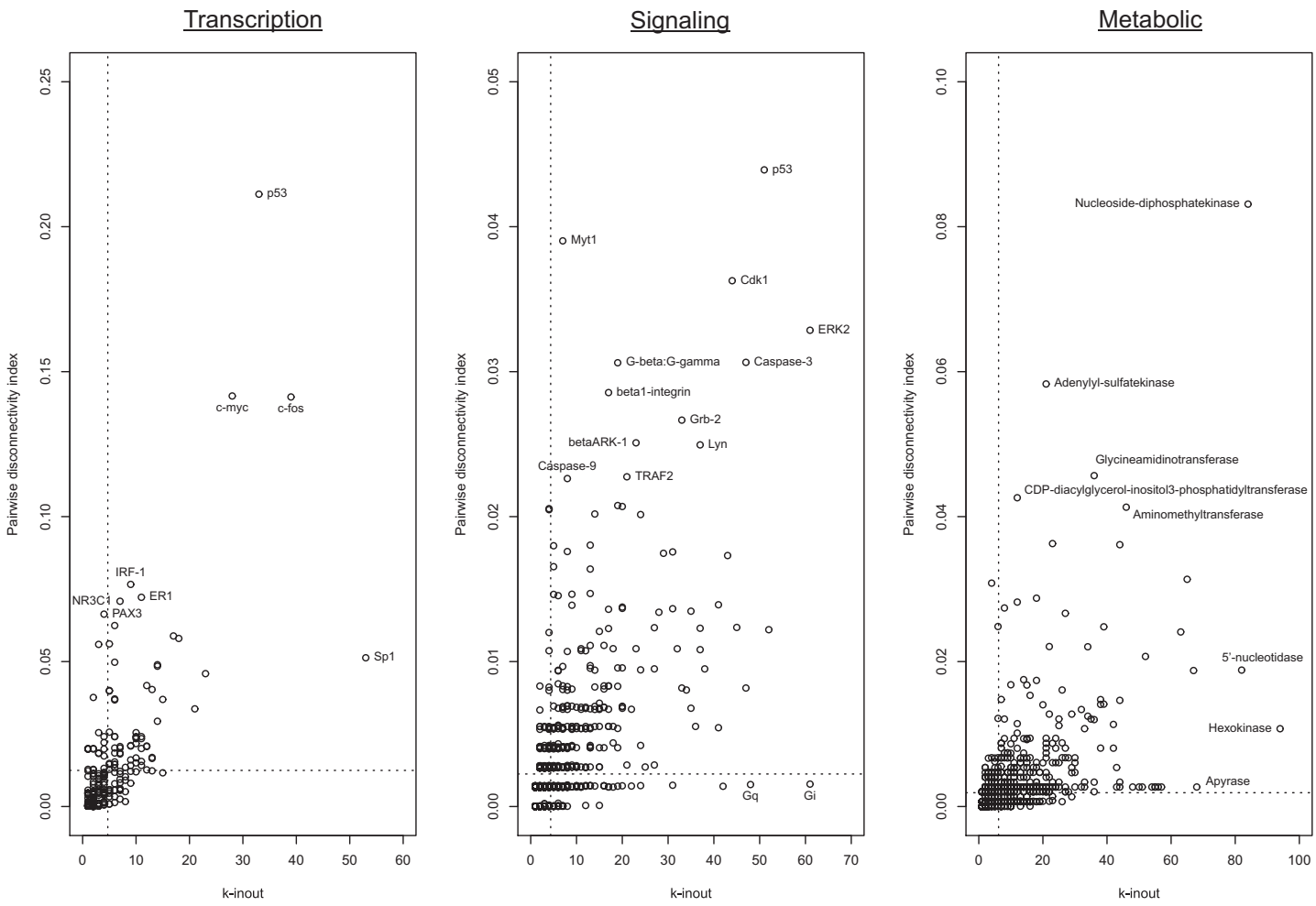

a)
Transcription
Signaling
Metabolic
C(k-inout)
k-inout
b)
log(C(k-inout))
log(k-inout)

Figure 4


Transcription
Signaling
Metabolic
Pairwise disconnectivity index
k-inout
p53
c-myc
c-fos
IRF-1
ER1
NR3C1
PAX3
Sp1
Myt1
Cdk1
ERK2
G-beta:G-gamma
Caspase-3
beta1-integrin
Grb-2
betaARK-1
Lyn
Caspase-9
TRAF2
Gq
Gi
Nucleoside-diphosphatekinase
Adenylyl-sulfatekinase
Glycineamidinotransferase
CDP-diacylglycerol-inositol3-phosphatidyltransferase
Aminomethyltransferase
5'-nucleotidase
Hexokinase
Apyrase

Additional files provided with this submission:

Additional file 1: Table1.pdf, 18K
http://www.biomedcentral.com/imedia/1694434521521914/supp1.pdf
Additional file 2: Table2.pdf, 21K
http://www.biomedcentral.com/imedia/6067264975219146/supp2.pdf
Additional file 3: Table3.pdf, 23K
http://www.biomedcentral.com/imedia/9217591185219147/supp3.pdf

Literatur

[1] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824 – 827, 2002.

[2] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551 – 1555, 2002.

[4] R. Albert, H. Jeong, and A. Barabási. Lethality and centrality in protein networks. *Nature*, 401:130 – 131, 1999.

[5] S. Maslov and K. Sneppen. Specifity and stability in topology of protein networks. *Science*, 296:910 – 913, 2002.

[6] A. Wagner and D.A. Fell. The small world inside large metabolic networks. *Proc R Soc Lond B Biol Sci*, 268:1803 – 1810, 2001.

[7] H. Ma and A.P. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19:270 – 277, 2003.

[8] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18:1283 – 1292, 2001.

[9] M. Newman. Assortative mixing in networks. *Phys Rev*, 89, 2002.

[10] M. Girvan and M. Newman. Community structure in social and biological networks. *PNAS*, 99:7821 – 7826, 2002.

[11] S.H. Yook, Z. Oltvai, and A. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928 – 942, 2004.

[12] J.A. Papin, T. Hunter, B.O. Palsson, and S. Subramaniam. Reconstruction of cellular signaling networks and analysis of their properties. *Nat Rev Mol Cell Biol*, 6:99 – 111, 2005.

[13] H. Agrawal. Extreme self-organization in networks constructed from gene expression data. *Phys Rev Lett*, 89:268702, 2002.

[14] A. Bhan, D.J. Galas, and T.G. Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18:1486 – 1493, 2002.

[15] N.M. Luscombe, M.M. Babu, H. Yu, M. Snyder, S.A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308 – 312, 2004.

[16] N. Guelzim, D. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31:60–63, 2002.

[17] R. Khanin and E. de Wit. How scale-free are biological networks? *Journal of Computational Biology*, 3:810–818, 2006.

[18] A.S. Konagurthu and A.M. Lesk. On the origin of distritbution patterns of motifs in biological networks. *BMC Systems Biology*, 2:73–81, 2008.

[19] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931 – 945, 2004.

[20] Mouse Genome Sequencing Consortium Asif T. Chinwalla et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520 – 562, 2002.

[21] Rat Genome Sequencing Project Consortium. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428:493 – 521, 2004.

[22] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. Kel, and E. Wingender. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acid Research*, 34:D108 – D110, 2006.

[23] M. Krull, S. Pistor, N. Voss, A. Kel, I. Reuter, D. Kroneberg, H. Michael, K. Schwarzer, A. Potapov, C. Choi, O. Kel-Margoulis, and E. Wingender. Transpath: An information resource for storing and visualizing signaling pathways and theor pathological aberrations. *Nucleic Acid Research*, 34:D546–D551, 2006.

[24] C. Choi, T. Crass, A. Kel, O. Kel-Margoulis, M. Krull, S. Pistor, A. Potapov, N. Voss, and E. Wingender. Consistent re-modeling of signaling pathways

and its implementation in the transpath database. *Genome Informatics Series*, 2:244 – 254, 2004.

[25] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acid Research*, 36:D480–D484, 2008.

[26] R. Albert, H. Jeong, and A. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[27] A. Barabási and Z. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews*, 5:101–113, 2004.

[28] A. Barabási. Scale-free networks: a decade and beyond. *Science*, 325:412–413, 2009.

[29] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley. Classes of small-world networks. *PNAS*, 97:11149–11152, 2000.

[30] A. Clauset and M.E.J. Shalizi, C .and Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

[31] L. Alderson, R. Tanaka, J. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2:431–523, 2005.

[32] J.M. Montoya, S.L. Pimm, and R.V. Solé. Ecological networks and their fragility. *Nature*, 442:259 – 263, 2006.

[33] B. Goemann, E. Wingender, and A. Potapov. The large-scale organization of mammalian networks with different functionallities: transcription, signal transduction and metabolic networks. *BMC Systems Biology*, 2011 (submitted).

[34] S A. Teichmann and M.M. Babu. Gene regulatory network growth by duplication. *Nature Genetics*, 36:492 – 496, 2004.

[35] S. Balaji, M.M. Babu, and L. Aravind. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of e. coli. *J Mol Biol*, 372:1108 – 1122, 2007.

[36] G. Lima-Mendez and J. van Helden. The powerful law of the power law and other myths in network biology. *Mol Biosyst*, 12:1482 – 1493, 2009.

[37] W.C. Liu, W.H. Lin, A.J. Davis, F. Jordán, H.T. Yang, and M.J. Hwang. A network perspective on the topological importance of enzymes and their phylogenetic conservation. *BMC Bioinformatics*, 8:121, 2007.

[38] A. Potapov, N. Voss, N. Sasse, and E. Wingender. Topology of mammalian transcription networks. *Genome informatics series*, 16:270 – 278, 2005.

[39] H. Yu, P.M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3:e59, 2007.

[40] P. Holme, M. Huss, and H. Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19:532 – 538, 2003.

[41] L.C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35 – 41, 1977.

[42] D. Koschützki and F. Schreiber. Comparison of centralities for biological networks. In *Proceedings of the German Conference on Bioinformatics (GCB 04) LNI P-53*, pages 199 – 206. Springer Verlag, 2004.

[43] B.H. Junker, D. Koschützki, and F. Schreiber. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, 7:219 – 225, 2006.

[44] A.P. Potapov, B. Goemann, and E. Wingender. The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks. *BMC Bioinformatics*, 9:227, 2008.

[45] J. Gross and J. Yellen. *Graph Theory and Its Applications*. CRC Press, 1998.

[46] B. Goemann, E. Wingender, and A. P. Potapov. An approach to evaluate the topological significance of motifs and other patterns in regulatory networks. *BMC Systems Biology*, 3:53, 2009.

[47] B. Goemann, T. Schoeps, A. Potapov, and E. Wingender. DiVa online: Finding key elements in regulatory networks by means of the pairwise disconnectivity index. *Bioinformatics*, 2011 (submitted).

[48] http://www.bioinf.med.uni-goettingen.de/services/diva.

[49] http://diva.sybig.de.

[50] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional network of e. coli. *Nature genetics*, 31:1 – 5, 2002.

[51] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:409–411, 1998.

[52] S. Akira and K. Takeda. Toll-like receptor signaling. *Nat Rev Immunol*, 4:499 – 511, 2004.

[53] A.P. West, A.A. Koblansky, and S. Ghosh. Recognition and signaling by toll-like receptors. *Annu Rev Cell Dev Biol*, 22:409 – 437, 2006.

[54] A. Kolb, S. Busby, H. Buc, S. Garges, and S. Adhya. Transcriptional regulation by camp and its receptor protein. *Annu Rev Biochem*, 62:749 – 795, 1993.

[55] A. Martinez-Antonio and J. Collado-Vides. Identifiying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*, 6:482 – 489, 2003.

[56] H. Kitano. Biological robustness. *Nature*, 5, 2004.

[57] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.

[58] V. Spirin and L.A. Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100:12123 – 12128, 2003.

[59] U. Alon. *An Introduction to Systems Biology*. Chapman & Hall/CRC, 2007.

[60] S. Wuchty, Z.N. Oltvai, and A.L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35:176 – 179, 2003.

[61] R. Dobrin, Q.K. Beg, A.L. Barabási, and Z.N. Oltvai. Aggregation of topological motifs in the escherichia coli transcriptional regulatory network. *BMC Bioinformatics*, 5:10, 2004.

[62] U. Alon. Network motifs: theory and experimental approaches. *Nature*, 8, 2007.

[63] F. Hayot and C. Jayaprakash. A feedforward loop motif in transcriptional regulation: induction and repression. *J Theor Biol*, 7:133–143, 2005.

[64] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 100:11980 – 11985, 2003.

[65] S. Mangan, A. Zaslaver, and U. Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol*, 334:197 – 204, 2003.

[66] L.A. Venkatesh, W. Hsing, K.E. Gibson, and T.J. Silhavy. Multiple feedback loops are key to a robust dynamic performance of tryptophan regulation in escherichia coli. *FEBS Lett*, 63:234 – 240, 2004.

[67] O. Brandmann, J.E. Ferrell, R. Li, and T. Meyer. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science*, 310:496–498, 2005.

[68] S.A. Ramsey, J.J. Smith, D. Orrell, M. Marelli, T.W. Petersen, H. de Atauri, P. nd Bolouri, and J.D. Aitchison. Dual feedback loops in the gal regulon suppress cellular heterogeneity in yeast. *Nature Genetics*, 38:1082–1087, 2006.

[69] D. Kim, Y.K. Kwon, and K.H. Cho. Coupled positive and negative feedback circuits form an essential building block of cellular signaling pathways. *BioEssays*, 29:85–90, 2007.

[70] J.R. Kim, Y. Yoon, and K.H. Cho. Coupled feedback loops form dynamic motifs of cellular networks. *Biophys J*, 94:359–365, 2008.

[71] Y. Artzy-Randrup, S.J. Fleishmann, N. Ben-Tal, and L. Stone. Comment on network motifs: Simple building blocks of complex networks and superfamilies of evolved and designed networks. *Science*, 20:1107, 2004.

[72] A. Mazurie, S. Bottani, and M. Vergassola. An evolutionary and functional assessment of regulatory network motifs. *Genome Biology*, 6:R35, 2005.

[73] P.J. Ingram, M.P.H. Stumpf, and J. Stark. Network motifs: structure does not determine function. *BMC Genomics*, 7:108, 2006.

[74] A. Vazquez, R. Dobrin, D. Sergi, J. Eckmann, Z.N. Oltvai, and A.L. Barabási. The topological relationship between the large-scale attributes and local interaction patterns in complex networks. *PNAS*, 101(52):17940 – 17945, 2001.

[75] S. Kalir, S. Mangan, and U. Alon. A coherent feed-forward loop with a sum input function prolongs flagella expression in escherichia coli. *Mol Sys Biol*, 1:2005 – 2006, 2005.

[76] B. Goemann, A. Potapov, M. Ante, and E. Wingender. Comparative analysis of topological patterns in different mammalian networks. *Genome Informatics Series*, 23, 2009.

[77] M. Castedo, J.L. Perfettini, T. Roumier, K. Andreau, R. Medema, and G. Kroemer. Cell death by mitotic catastrophe: a molecular definition. *Oncogene*, 23:2825–2837, 2004.

[78] A. Wagner. Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol.*, 19:1760 – 1768, 2002.

[79] L. Danon, A. Arenas, and A. Díaz-Guilera. Impact of community structure on information transfer. *Phys Rev E*, 77:036103, 2008.

Lebenslauf

Persönliche Informationen

Name	Björn Goemann
Geburtsdatum	09.08.1975
Geburtsort	Einbeck
Staatsangehörigkeit	Deutsch

Schulbildung

1992 - 1995	Gymnasium der Berufsbildenden Schulen, Einbeck

Zivildienst

1995 - 1996	Jugendgästehaus des Deutschen Roten Kreuzes, Einbeck

Berufsausbildung

1996 - 1999	Ausbildung zum Datenverabeitungskaufmann bei der BSV Software GmbH, Göttingen

Studium

1999 - 2005	Diplomstudiengang Wirtschaftsinformatik an der Georg-August Universität Göttingen Diplomarbeit: „Konzeption und Implementierung eines Systems zur Darstellung unscharfer Kennzahlen“

Beruflicher Werdegang

1999	IT Consultant bei der BSV Software GmbH, Göttingen
2000 - 2002	Softwareentwickler im Rahmen einer studentischen Nebentätigkeit bei der Prof. Schumann GmbH, Göttingen
2002 - 2003	Softwareentwickler im Rahmen einer studentischen Nebentätigkeit bei der B&N Crossgate AG, Göttingen
2005	Wissenschaftliche Hilfskraft der Abteilung Bioinformatik des Universitätsklinikums Göttingen
2006 - heute	Wissenschaftlicher Mitarbeiter der Abteilung Bioinformatik des Universitätsklinikums Göttingen