Eberhard Karls Universität Tübingen
Fakultät für Informations- und Kognitionswissenschaften
Wilhelm-Schickard-Institut
Arbeitsbereich: Rechnerarchitektur

Seminar
Inferenz regulatorischer und metabolischer Netze
in der Systembiologie

Thema: Inferenz genregulatorischer Netzwerke

Andreas Jahn

Zusammenfassung: ***Die Konstruktion genregulatorischer Netzwerke, die funktionale Zusammenhänge korrekt wiedergeben, sind für die Biologie von großem Interesse. Die Informationen, die aus solchen Netzwerken gewonnen werden können, liefern wichtige Hinweise, die das Verständnis eines breiten Spektrums anderer biologischer Probleme förden könnte. Jedoch erschwert die große Anzahl an regulatorischen Beziehungen ein intuitives Verständnis und fordert den Einsatz von computergestützten Modellen. Die grundlegenden Prinzipien der Inferenz soll in dieser Arbeit anhand von drei verschiedenen Modellen erklärt werden.***

Inhaltsverzeichnis

1 Einleitung & Motivation

Komplexe Organismen bestehen aus einer Vielzahl unterschiedlicher Zellen, die sich im Laufe der Evolution an spezifische Aufgaben angepasst haben. Dabei kann man die einzelnen Zelltypen nicht nur makroskopisch, sondern auch biochemisch unterscheiden. Alle Zellen eines Organismusses enthalten die gleichen genetischen Informationen, können sich aber deutlich in der Zusammensetzung der Proteinkonzentrationen unterscheiden (Friedman u. a. 2000). Die Proteinsynthese ist ein vielschichtiger Prozess, dessen Regulation auf mehreren Ebenen stattfindet. Ein elemtarer Mechanismus, der erstmals von Jacob und Monod beschrieben wird (Jacob u. Monod 1961), beruht auf einem komplexen Zusammenspiel von Transkriptionsfaktoren und den passenden *cis*-regulatorischen Elementen, welche sich in der Promoterregion eines Gens befinden (Nguyen u. D'Haeseleer 2006). Da diese Transkriptionsfaktoren selbst Proteine sind, kann der regulatorische Prozess als eine Summe von regulatorischen Verbindungen der einzelnen Gene betrachtet werden (Abbildung 1).

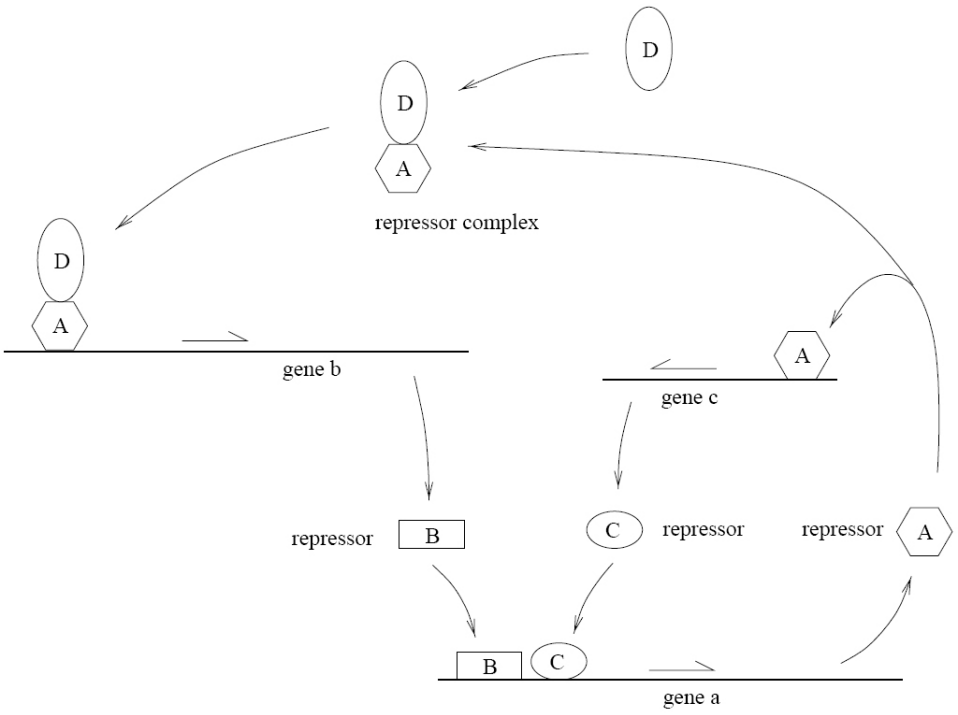

Abbildung 1: Beispiel eines regulatorischen Systems, bestehend aus drei Genen *a, b* und *c*, welche die Repressoren *A, B* und *C* kodieren. Die Repressorproteine inhibieren die Transkription der Gene die sie regulieren (de Jong 2002).

Die Identifizierung solcher regulatorischer Zusammenhänge ist wichtig für das Verständnis zellulärer Prozesse und Antworten auf interne und externe Signale (Segal u. a. 2003).

Eine große Zahl der regulatorischen Systeme bestehen aus vielen sich verzahnenden Elemente. So kann ein Transkriptionsfaktor die Regulation mehrere andere Gene beeinflussen und bei mehreren Tausend Genen ist es schwierig ein intuitives Verständnis des gesamten Netzwerkes zu erhalten. Der Einsatz von Computermodellen ermöglicht eine systematische Vorgehensweise und legt so den Grundstein auch das Verhalten von diffizilen Netzwerken zu erfassen.
In den letzten 30 Jahren wurden viele Modelle vorgeschlagen, die auf unterschiedlichen Verfahren wie, Booleschen Netzwerken (Kauffman 1974), Differenzialgleichungen (Chen u. a. 1999), Bayesschen Netzwerken (Friedman u. a. 2000) und regelbasierten Formalismen (Soinov u. a. 2003) beruhen. Diese Arbeit soll die Prinzipien der Inferenz genregulatorischer Netzwerke anhand von drei Modellen veranschaulichen. Das erste Modell stellt einen der ersten Ansätze dar. Der zweite und dritte Ansatz sind aktuelle Arbeiten, die auf unterschiedlichen Modellen beruhen.

1.1 Netzwerke als Graphen

Der intuitivste Ansatz ein genregulatorisches Netzwerk zu repräsentieren, ist die Darstellung als Graph. Ein gerichteter Graph *G* ist durch ein Tupel (V, E) definiert, wobei *V* eine Menge von Knoten und *E* eine Menge von Kanten symbolisieren. Die Knoten eines gerichteten Graphen stellen Gene oder andere Elemente des Netzwerkes dar. Die Definition der Kanten kann zwischen den einzelnen Modellen stark variieren. Es kann aber generell gesagt werden, dass eine Kante (i, j, k) einen Zusammenhang zwischen den Genen *i* und *j* beschreibt. Der Informationsgehalt von *k* ist abhängig von dem zu Grunde liegendem Modell und beschreibt die Interaktion der Gene *i* und *j*. Mögliche Interaktionen, die zwischen zwei Genen stattfinden können, sind zum Beispiel Aktivierungen und Hemmungen.
Die Regulation eines Gens kann natürlich auch von mehreren anderen Genen mit unterschiedlichen Interaktionen beeinflusst werden. Dies kann durch den Einsatz von Hypergraphen modelliert werden, bei dem die Elemente einer Kante auch aus Mengen bestehen können. Eine Kante (i, J, K) bedeutet, dass das Gen *i*, von einer Menge an Genen *J* mit den Interaktionen *K*, reguliert wird.

2 Boolesche Netzwerke

2.1 Einleitung

Boolesche Netzwerke stellen eine radikale Idealisierung der Elemente eines regulatorischen Netzwerkes und deren Interaktionen dar (de Jong 2002). Der Zustand eines Gens wird durch eine Boolesche Variable beschrieben und kann dadurch nur zwei Zustände einnehmen. Interaktionen zwischen einzelnen Elementen werden als Boolesche Funktio-

nen dargestellt und können somit berechnet werden.
Der Zustand eines Booleschen Netzwerkes mit *n* regulatorischen Elementen kann durch *n* Variablen festgelegt werden. Jede Variable besitzt einen Wert von *0* oder *1*, wodurch das System 2^n mögliche Zustände einnehmen kann. Für jede der *n* Variablen des Systems existiert eine Boolesche Funktion, deren Auswertung den Folgezustand der Variable festlegt. Diese Zustandsänderungen sind deterministisch, wodurch für eine gegebene Eingabe eine eindeutig definierte Ausgabe existiert. Die Aktualisierung der einzelnen Ausgänge erfolgt in diskreten Zeitschritten und synchron.
Die Idee genregulatorische Netzwerke durch Boolesche Netzwerke zu modellieren stammt aus dem Jahr 1969 (Kauffman 1969). Der Ansatz und die grundliegende Funktion soll anhand einer späteren Arbeit von Kauffman dargestellt werden (Kauffman 1974).

2.2 Das Modell

Stuart Kauffman betrachtete im Jahr 1973 das Lac-Operon von *Escherichia coli* und dessen Regulation durch den Operatorzustand. Der Operator, der in seinem gebundenen Zustand die Transkription der Gene *Z, Y, A* verhindert, wird durch zwei Moleküle bestimmt (Kauffman 1974). Einem Repressormolekül, welches durch das Regulatorgen *I* kodiert wird und einem Derivat der Laktose, die Allolaktose, welche an den Repressor bindet und somit die Bindungsaffinität zwischen Operator und Repressor reduziert.
Kauffman erkannte, dass der Zustand des Operators durch die Konzentrationen der Kontrollmoleküle festgelegt ist. Der Operator kann nur gebunden sein, wenn der Repressor in einer ausreichenden Konzentration vorhanden ist und sich keine Allolaktose in der Zelle befindet. Bei einer ausreichender Allolaktosekonzentration, oder Fehlen von Repressormolekülen, ist das Repressormolekül in einem ungebundenen Zustand. Anhand dieser Informationen lässt sich der Zustand des Operators durch eine Boolesche Funktion ausdrücken. Sei der Zustand des Operators definiert durch *Op*, wobei *Op=1* bedeutet, dass die Allolaktose an den Operator gebunden ist. Die Allolaktose und der Repressor werden jeweils durch *Allolaktose* und *Repressor* symbolisiert, mit den möglichen Zuständen *Allolaktose=1* bzw. *Repressor=1*, wenn das jeweilige Molekül in einer ausreichenden Konzentration vorhanden ist. Der Zustand des Operators kann durch diesen Formalismus einfach in einer Tabelle mit allen möglichen Kombinationen beschrieben werden. Diese Tabelle kann als Boolesche Funktion interpretiert werden (Tabelle 1).
Die Tabelle zeigt auch, dass der Zustand der Booleschen Funktion nicht immer von allen Eingangsvariablen abhängig ist. Für das Beispiel des Lac-Operons definiert die Eingangsvariable *Repressor=0* bereits eindeutig *Op=0*, unabhängig von der zweiten Eingangsvariable. Genauso trifft diese Aussage für *Allolaktose=1* zu. Nur für den Zustand *Op=1* benötigt es einer Koordination beider Variablen. Solch eine Funktion, bei der mindestens eine Eingangsvariable einen Zustand besitzt, bei dem die Variable alleine den Zustand der Funktion bestimmt, wird auch Kanalfunktion genannt (Kauffman 1974). Solche booleschen Kanalfunktionen können auch in anderen Organismen, wie dem Bakteriophagen λ,

Allolaktose	Repressor	Op
0	0	0
0	1	1
1	0	0
1	1	0

Tabelle 1: Die Tabelle zeigt alle Kombinationsmöglichkeiten der beiden Eingangsvariablen und den daraus resultieren Zustand für den Operator.

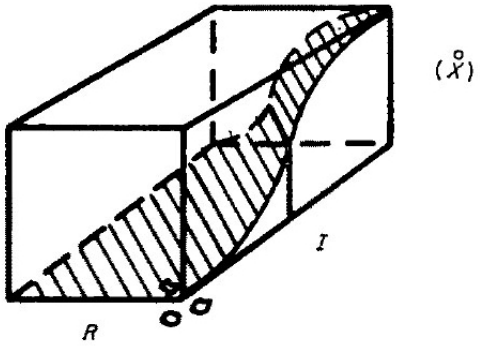

Abbildung 2: Eine mögliche Darstellung einer kontinuierlichen und monotonen Bindungsfunktion für die Aktivität X eines biologischen Prozesses als Funktion von den Konzentrationen des *Inducer (I)* und *Repressor (R)* (Kauffman 1974).

identifiziert werden (Kauffman 1974).

2.3 Fazit

Biologische Prozesse, wie Genaktivität, zeigen ein kontinuierliches Verhalten und sollten besser durch Michaelis-Menten oder kooperativen sigmoidalen Bindungsfunktionen (Abbildung 2) beschrieben werden (Kauffman 1974). Boolesche Funktionen stellen dagegen eine Idealisierung dar, die einerseits gut in ihrer Handhabung sind, aber andererseits die Zustände der Gene auf zwei diskrete Werte beschränken. In diesem Zusammenhang gibt es gute Gründe andere Modelle zu verwenden. Trotzdem werden biologische Vorgänge oftmals durch eine binär-logische Sprache, wie “an” und “aus” oder “hochreguliert” und “runterreguliert”, beschrieben. Es existieren eine Reihe von Beispielen die zeigen, dass ein Boolescher Formalismus durchaus aussagekräftig sein kann und deren Einsatz rechtfertigt (Wosik 2004; Zhang u. Shmulevich 2002).

3 Regelbasiertes Modell

3.1 Einleitung

Regelbasierte Modelle stammen aus dem Gebiet der künstlichen Intelligenz und stellen eine weitere Möglichkeit dar genregulatorische Netzwerke zu beschreiben. Sie versuchen anhand von Informationen, wie Expressionsmatrizen, Zusammenhänge zwischen den einzelnen Genen zu erkennen und den Expressionsgrad eines Gens durch Betrachtung anderer Gene vorherzusagen. Diese Vorhersagen werden anhand klarer Regeln erstellt, welche durch Entscheidungsbäume oder durch eine Grammatik definiert sind (Abbildung 3 und Tabelle 2). Diese Methode benötigt keine *a priori* Diskretisierung der Daten und funktioniert auch mit kontinuierlichen Expressionsdaten (Soinov u. a. 2003). Zusätzlich erlaubt dieser Ansatz die Definition von verschiedenen Grenzwerten eines regulierenden Gens, so dass ein Gen mehrere andere Gene bei verschiedenen Konzentrationen beeinflusst. Dadurch unterscheidet sich diese Methode von den Booleschen Netzwerken, bei denen die Eingangsvariablen *a priori* diskretisierte Werte haben und diese für das gesamte System gelten. Das Verfahren solcher Modelle soll exemplarisch anhand einer Arbeit von Soinov u. a. dargestellt werden.

3.2 Das Modell

Das Modell beruht auf einem überwachten maschinellen Lernverfahren und versucht durch *Classifier* den Zustand eines Gens vorherzusagen. Die Vorteile die durch den Einsatz von *Classifiern* entstehen, sind einerseits die Transparenz der *Classifier*, die es ermöglichen Zusammenhänge zwischen den einzelnen Genen direkt den *Classifiern* zu entnehmen, andererseits können diese einfach in Form von Entscheidungsbäumen und Regeln dargestellt werden und sind somit einfach zu interpretieren (Soinov u. a. 2003).
Die Transkriptionseinheit, die den Expressionsgrad eines Gens bestimmt, wird in diesem Modell in eine endliche Anzahl diskreter Zustände eingeteilt. Auf Grund der einfacheren Darstellung des Modells, werden nur zwei Zustände definiert. Die beiden Zusände kodieren die Aussagen, ob sich der Expressionsgrad eines Gens “über dem Durchschnitt” oder “unter dem Durchschnitt” befindet.
Das Problem, mit welchem das Modell konfrontiert wird, lässt sich wie folgt beschreiben. Ausgehend von der Expressionsmatrix X und einem Gen i kann man nun drei Teilprobleme betrachten.

1. Der Zustand eines Gens i in Probe j wird von den Expressionswerten der anderen Gene der Probe vorhergesagt.

2. Der Zustand eines Gens i in Probe j wird von den Expressionswerten vorheriger Proben bestimmt.

3. Die Zustandsänderung des Gens *i* wird durch die Zustandsänderungen anderer Gene ermittelt.

Für eine formelle Definition dieser Probleme müssen zunächst einige Definitionen eingeführt werden. Sei die Spalte $y_j = (x_{1j}, \ldots, x_{kj})$ der Genexpressionsmatrix *X* das Expressionsmuster einer Probe. Mit $y_{j/i}$ sei nun ein Expressionsmuster definiert, bei dem der Expressionswert für das Gen *i* fehlt. Der Zustand der Transkriptionseinheit von Gen *i* in Probe *j* sei s_{ij} und wie in Gleichung 1 definiert. $Y = \{y_1, \ldots, y_n\}$ sei die Menge aller Expressionsmuster und $Y_{/i} = \{y_{1/i}, \ldots, y_{n/i}\}$ analog dazu die Menge aller Expressionsmuster, bei denen der Expressionswert für das Gen *i* fehlt.
Ein *Classifier C* stellt nun eine Funktion dar, die einen Vektor *y* auf diskrete Werte *s* abbildet. Die Teilmenge von *y*, welcher korrekte Werte zugewiesen werden können, nennt man Datensatz *D*. Ein Induktionsalgorithmus ist ein Verfahren, das den Datensatz *D* auf einen *Classifier C* abbildet. Für eine Lösung der obengenannten Probleme müssen die Datensätze definiert und einen geeigneten Induktionsalgorithmus gewählt werden. Die Bildung der Datensätze kann anhand der eingeführten Definitionen präziser formuliert werden.

$$s_{ij} = \begin{cases} +1, & \text{wenn } x_{ij} > \overline{x_i}, \\ & \text{mit } \overline{x_i} \text{ dem durchschnittlichem Expressionsgrad von Gen } i \\ -1, & \text{sonst} \end{cases} \quad (1)$$

Bei dem ersten Problem soll der Zustand des Gens *i* durch das Expressionsmuster der anderen Genen der Probe bestimmt werden. Der Induktionsalgorithmus *I* bildet hier den Datensatz $D^i = (Y_{/i}, s_i)$ auf den *Classifier* C^i ab. Für den gegebenen Datensatz muss also ein *Classifier* konstruiert werden, der dem Gen *i* den korrekten Zustand zuweist. Es muss also die Gleichung $I(D^i, y_{j/i}) = C^i(y_{j/i}) = s_{ij}$ erfüllt sein.
Die genaue Definition des zweiten Problems ist ähnlicher Natur mit der Ausnahme, dass der Datensatz durch $D^i = (Y'_{/i}, s'_{/i})$ festgelegt ist, mit $Y'_i = \{y_{1/i}, \ldots, y_{n-1/i}\}$ und $s'_i = (s_{i2}, \ldots, s_{in})$. Eine korrekte Klassifizierung durch den *Classifier* ist erreicht, wenn die Gleichung $C^i(y_{j/i}) = s_{ij+1}$ gilt.
Für das dritte Problem muss zunächst in einem ersten Schritt eine Matrix *D* erstellt werden, die aus den Elementen $d_{ij} = s_{ij+1} - s_{ij}$ besteht. Die einzelnen Elemente der Matrix können durch die Restriktion der Zustandsmenge der Transkriptionseinheit nur Werte aus der Menge $\{-1, 0, +1\}$ annehmen. Ein Wert von *+1* bedeutet eine Änderung des Zustandes der jeweiligen Transkriptionseinheit von "runterreguliert" zu "hochreguliert", *-1* die Gegenteilige Änderung und *0* symbolisiert keine Veränderung des Zustandes. Die Lösung dieses Problems kann auf die des ersten Problems zurückgeführt werden, mit der Änderung, dass die Einträge der Matrix *D* anstatt der Werte der Expressionsmatrix *X* verwendet werden.
Ein weiteres Problem wirft die Frage auf, welche der Gene aus einem Datensatz *D* sind wirklich relevant für eine korrekte Vorhersage, oder biologisch formuliert: welche Ge-

ne haben einen regulatorischen Einfluss auf das betrachtete Gen *i*. Dieses Problem kann durch *feature subset selection* angegangen werden und es existieren zwei verbreitete Methoden dieses zu lösen. Die *Filter* Methode filtert einzelne Komponenten aus und versucht die beste Kombination durch eine Bewertung einer Zielfunktion zu finden. Diese Methode bezieht keine Informationen des Induktionsalgorithmusses mit ein, dies kann als Schwachpunkt des Verfahrens angesehen werden. Im Gegensatz dazu versucht die *Wrapper* Methode die Parameter des Induktionsalgorithmusses zu verbessern und dadurch eine Optimierung der Zielfunktionen zu erreichen.

Die allgemeine Vorgehensweise, bei der Erstellung eines genregulatorischen Netzwerkes mit dem beschriebenen Modell, sieht wie folgt aus. In einem ersten Schritt muss das Modell anhand eines Datensatzes trainiert werden. In diesem Beispiel wurden die Microarraydaten von Spellman u. a. und Cho u. a. verwendet (Spellman u. a. 1998; Cho u. a. 1998). Diese Datensätze enthalten Messungen des Zellzyklusses der *Saccharomyces cerevisiae*. Das überwachte Training wurde mit dem *cdc15* Datensatz durchgeführt, da dieser die meisten Messwerte enthält. Die Genauigkeit der Ergebnisse des Modells wurde durch drei Verfahren bestimmt. Einer zehnfachen Kreuzvalidierung und die Verwendung der Datensätze *cdc28* und *alpha-factor* als Testdatensätze.
Der zweite Schritt besteht darin, die Informationen der *Classifier* zu extrahieren und mit biologischen Daten zu überprüfen. Eine Überprüfung mittels biologischen Daten stellt einen wichtigen Punkt dar, da eine Validierung des Modells nur anhand der Vorhersagegenauigkeit nichts über die Korrektheit der biologischen Interpretation der Daten aussagt. Damit ein solcher Vergleich überhaupt erst realisiert werden kann, müssen die Informationen aus den *Classifiern* in kompakter und einfacher Form dargestellt werden können. Eine einfache Regelsprache genügt dieser Form und kann wie folgt definiert werden. Alle Gene werden mit Großbuchstaben versehen und ein Vorzeichen gibt Auskunft über den Expressionsgrad. So bedeutet “$+A$”, dass das Gen *A* hochreguliert ist. Interaktionen zwischen Genen werden durch Pfeile, wie $\Rightarrow$ und $\Leftrightarrow$ symbolisiert.
Mit den so gewonnen Informationen kann ein genregulatorisches Netzwerk, wie es in der Abbildung 4 dargestellt ist, erstellt werden.

3.3 Ergebnisse

Ziel des Modells ist es anhand von Microarraydaten genregulatorische Zusammenhänge zwischen den Genen zu entdecken. Da es sich bei den Daten um Messungen der Expressionswerte der einzelnen Stadien des Zellzyklusses handelt, wird besonderen Wert auf die Gene gelegt, von denen bekannt ist, dass sie eine signifikante Rolle in der Regulation des Zellzyklusses spielen. Das Modell arbeitet mit zwei möglichen Zuständen für jede Transkriptionseinheit, so dass nur Regeln aus den *Classifiern* übernommen werden können, die durch zwei Zustände eindeutig definiert sind.
Die Tabelle 2 zeigt eine Auflistung einiger Regeln für die Gene, die bei der Regulation

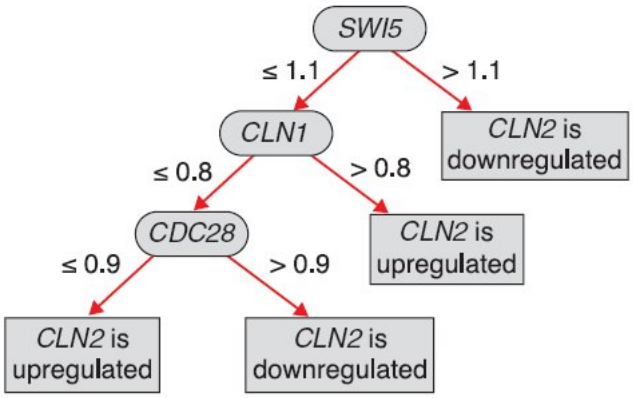

Abbildung 3: Die Abbildung zeigt einen Entscheidungsbaum für das Zielgen *CLN2*. Die Gene *SWI5, CLN1* und *CDC28* stellen die erklärenden Gene dar. Expressionsschranken für die entsprechenden erklärenden Gene stehen an allen Ästen (Soinov u. a. 2003).

$-CLB1 \Leftrightarrow -SWI5$	$-CLB2 \Leftrightarrow -SWI5$
$+CLN2 \Leftrightarrow +CLN1$	$-CDC20 \Leftrightarrow +CLN1$
$-CLB2 \Leftrightarrow +CLN2$	$+SWI5 \Leftrightarrow -CLN2$
$\pm CLB2 \Leftrightarrow \pm CLB1$	$-CLB1 \Leftrightarrow -CLB2$

Tabelle 2: Die Tabelle zeigt eine Auswahl der Regeln, die aus den *Classifiern* extrahiert wurden. Es sind nur Regeln aufgelistet, die bei allen drei Testverfahren eine hohe Genauigkeit erzielten.

des Zellzyklusses eine Rolle spielen. Alle diese Regeln können durch publizierte Arbeiten bestätigt werden und können somit als korrekt angesehen werden. Ausgehend von diesen Regeln ist die Erstellung eines genregulatorischen Netzwerkes eine einfache Aufgabe und führt zu einem Graphen, wie er in Abbildung 4 zu sehen ist.

3.4 Fazit

Obwohl das Modell den Expressionsgrad eines Gens nur durch zwei diskrete Werte kodiert, zeigen alle gewonnen Regeln biologische Relevanz und können durch die Literatur bestätigt werden. Die Verwendung von *Classifiern* ermöglichen eine einfache Informationsgewinnung und führen zu klar definierten Regeln. Dies simplifiziert wiederum die Validierung der Regeln mit biologischen Daten. Ein genregulatorisches Netzwerk, welches anhand der Regeln erstellt wird, ist eindeutig und stellt die Informationen klar dar.

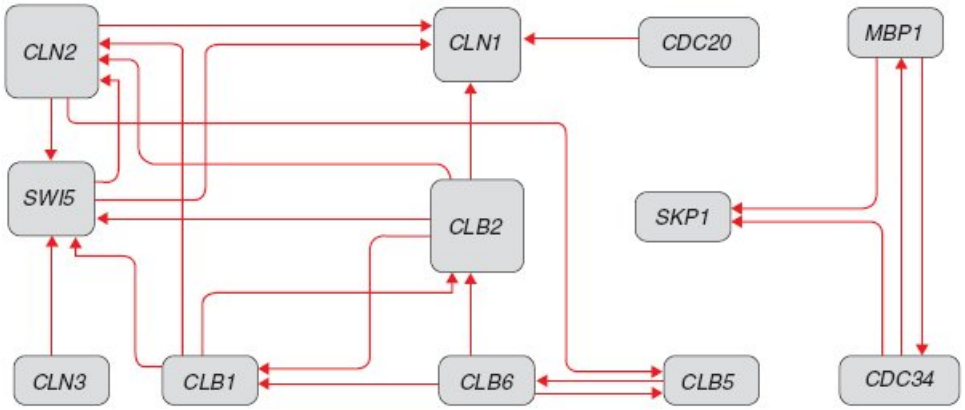

Abbildung 4: Die Abbildung zeigt das genregulatorische Netzwerk, das durch die Entscheidungsregeln des *cdc15* Datensatzes erstellt wurden. Es stellt eine graphische Repräsentation der Informationen dar, die aus den Regeln abgeleitet sind. Die Knoten des Graphen symbolisieren die Gene und die Kanten die Beziehung zwischen den Genen. Zusätzlich kann man der Grafik entnehmen, dass zwei Module in dem Netzwerk existieren (Soinov u. a. 2003).

Kritisch muss dagegen die Reduzierung der Gene auf eine Teilmenge, die bei dem Zellzyklus regulatorische Funktionen übernehmen, gesehen werden. Die Darstellung des Expressionsgrades eines Gens durch lediglich zwei Zustände hat zur Folge, dass es bei der Erstellung der Regeln zu einer starken Kompression der Daten kommt, dies kann zu einem Informationsverlust führen. Dieser Verlust kann zur Folge haben, dass wichtige Interaktionen, die bei einer geringen Zustandsmenge nicht aufgelöst werden können, verloren gehen und somit das erstellte genregulatorische Netzwerk nicht vollständig ist.
Das Modell kann noch durch zwei unterschiedliche Strategien verbessert werden. Eine Möglichkeit stellt die Erstellung größerer Datensätze mit mehr Messpunkten dar. Ein Training des Modells an solchen Datensätzen kann zu einer signifikanten Verbesserung führen. Zusätzlich sollte die Anzahl der Zustände, die eine Transkriptionseinheit einnehmen kann, erhöht werden. Durch eine feinere Auflösung können auch komplexere Zusammenhänge zwischen den einzelnen Genen erkannt werden, dies steigert den Informationsgewinn. Die Steigerung der Zustandsmenge kann aber ab einer gewissen größe auch das Gegenteil bewirken. Das Modell erstellt dann sehr komplexe Regeln, die von vielen Genen abhängen. Ein genregulatorisches Netzwerk, das durch solche komplexen Regeln erstellt wird, wäre ein beinahe vollständig verbundener Graph und würde keine biologisch relevanten Informationen enthalten.

4 Die MED-Methode

4.1 Einleitung

Die *Matrix-Expressions-Dekompositions*-Methode (MED-Methode), die zu den mathematisch-linearen Modellen gehört, hebt sich von den bereits vorgestellten Modellen deutlich ab. Im Gegensatz zu den anderen Modellen versucht dieses Modell keine Vorhersage eines Expressionsgrades aufgrund der Expressionsgrade anderer Gene zu treffen, sondern verfolgt eine andere Strategie. Mit dem biologischen Hintergrundwissen, dass die Expression eines Gens zu einem großen Teil von der Beschaffenheit der Promotorregion und der darin enthaltenen *cis*-regulatorischen Elemente abhängt, liegt der Verdacht nahe, dass es möglich sein muss einen Zusammenhang zwischen den Eigenschaften der *cis*-regulatorischen Elemente und dem Expressionsgrad eines Gens herzustellen.
Grundlegende Voraussetzung für diese Methode ist eine möglichst komplette Sammlung aller *cis*-regulatorischer Elemente. Ein Verfahren, das die benötigten Informationen liefert, stammt von Kellis u. a. und beruht auf der Tatsache, dass wichtige Teile der DNS nur selten Mutationen unterliegen und daher konserviert sind (Kellis u. a. 2003). Dieses alternative Verfahren soll anhand einer Arbeit von Nguyen und D'haeseleer veranschaulicht werden (Nguyen u. D'Haeseleer 2006).

4.2 Das Modell

Die MED-Methode lässt sich allgemein in zwei Teilschritte aufteilen. In einem ersten Schritt wird jedes Gen einzeln untersucht und für jedes Motiv, das sich in der Promotorregion des Gens befindet, wird eine Motivstärke berechnet. Die Eigenschaften der Motive, wie Position und Kombinationen mit anderen Motiven, werden nicht bei der Berechnung berücksichtigt und spielen erst in dem zweiten Teilschritt eine Rolle. Die Idee, die hinter dieser Berechnung steht, beruht auf einem Modell von Jacob und Monod das besagt, dass der Expressionsgrad eines Gens, als Funktion der Motivmenge und den regulatorischen Faktoren der Zellumgebung definiert werden kann (Jacob u. Monod 1961). Nach diesem Modell gilt für den Expressiongrad *E* eines Gens *g* der in Gleichung 2 beschriebene Zusammenhang.

$$E_{gc} \approx \sum_{j \in \Omega_g} M_{gj} A_{jc} \tag{2}$$

M_{gj} stellt die Motivstärke des *j*-ten Motives in Gen *g* dar. A_{jc} repräsentiert eine allgemeine Näherungsvariable für das Motiv *j* unter den regulatorischen Faktoren der Zellumgebung *c*. Für alle Gene und Umgebungsbedingungen folgt aus Gleichung 2 die Gleichung 3.

$$E \approx M \bullet A \tag{3}$$

In Gleichung 3 stellt *E* eine $m \times n$ Matrix mit den Expressionswerten dar, wobei *m* die einzelnen Gene und *n* die unterschiedlichen Bedingungen unter denen gemessen wurde repräsentieren. *M* ist eine $m \times k$ Matrix mit *m* Genen und *k* Bedingungen und enthält die von der Umgebungsbedingungen unabhängigen Motivstärken. Wichtig ist, dass für ein Motiv *j*, welches nicht in der Promotorregion von Gen *i* liegt, das Element M_{ij} der Matrix *M* mit dem Wert *0* initialisiert wird. *A* ist eine $k \times n$ Matrix mit den Näherungsfaktoren für die Motive *k* unter den Umgebungsbedingungen *n*.
Die interessanten Informationen, die Motivstärken, sind in der Matrix *M* enthalten. Die Methode muss also ausgehend von den Expressionsdaten und den Informationen der Promotorregionen eine eindeutige Matrix *M* berechnen. Dieses Problem wird durch eine Matrixdekomposition der Matrix *E* in ein Produkt der Matrizen *M* und *A* gelöst. Die Matrixdekomposition stellt einen iterativen Prozess dar, der deterministisch ist und eine eindeutige Lösung der Matrix *M* berechnet. Für den ausführlichen Beweis sei hier auf die Arbeit von Nguyen und D'haeseleer verwiesen (Nguyen u. D'Haeseleer 2006). Damit dieses Verfahren in der Lage ist eine eindeutige Lösung zu finden, muss die Matrix *M* vor dem iterativen Prozess mit geeigneten Werten initialisiert werden. Für jedes Motiv *m*, das in der Promotorregion von Gen *g* vorkommt, wird der Initialwert mit der Gleichung 4 berechnet.

$$M_{gm} = \sum_i \frac{e^{S_i(g,m)}}{\max_{i,g}(e^{S_i(g,m)})} \tag{4}$$

$S_i(g, m)$ ist die *ScanAce* Punktzahl der *i*-ten Instanz des Motives *m* in der Promotorregion von Gen *g* (Hughes u. a. 2000). Der anschließende iterative Prozess lässt sich durch drei Teilschritte beschreiben. Zuerst muss eine Matrix *A* erstellt werden, die die Summe aus Gleichung 5 minimiert.

$$\sum_{j=1}^{m} (E_{jc} - E'_{jc})^2 \tag{5}$$

Danach folgt eine Normalisierung der Matrix *A* in der Form, dass jede Zeile die Einheitsnorm besitzt. Ausgehend von den Matrizen *E* und *A* folgt zum Schluss eine Optimierung der Matrix *M*. Es wird eine optimale Motivstärke M_{gj} für jedes Motiv *j* in der Promotorregion von Gen *g* gesucht, die die Summe aus Gleichung 6 minimiert. λ stellt hierbei einen Faktor dar, der das Konvergenzkriterium festlegt.

$$\sum_{i=1}^{n} \left[E_{gi} - \sum_{j \in \Omega_g} M_{gj} A_{ji} \right]^2 + \lambda \sum_{j \in \Omega_g} M_{gj}^2 \tag{6}$$

Dieser letzte Schritt kann auch modifiziert werden, damit auch *a priori* bekannte Motivstärken vordefiniert werden können. Diese Modifizierung führt zu einer Optimierung anhand der Summe 7, wobei M^*_{gj} die vordefinierte Motivstärke für das Motiv *j* in der Promotorregion von Gen *g* darstellt. Dieser mehrstufige Prozess wird iterativ durchgeführt,

bis das Konvergenzkriterium erfüllt ist.

$$\sum_{i=1}^{n}\left[E_{gi}-\sum_{j\in\Omega_g}M_{gj}A_{ji}\right]^2+\lambda\sum_{j\in\Omega_g}\left[M_{gj}-M^*_{gj}\right]^2 \tag{7}$$

Der zweite Teilschritt der Methode besteht darin regulatorische Prinzipien aus der Matrix *M* abzuleiten. Hierfür wird zunächst ein *Genensemble* für eine bestimmte Motivmenge erstellt. Dieses *Ensemble* wird anschließend in einzelne Instanzen eingeteilt, in denen die Gene enthalten sind, die bestimmten Anforderungen an die Eigenschaften der Motive genügen (Abbildung 5). Die Anforderungen an die Eigenschaften der Motive kann unterschiedlicher Natur sein und muss nicht nur auf die Position der Motive in der Promotorregion beschränkt sein. Für alle erstellten Instanzen eines *Genensemble* lässt sich anhand der Matrix *M* aus dem ersten Teilschritt, die durchschnittliche Motivstärke und der Standardfehler berechnen.

Für die Ergebnisse, die in Abschnitt 4.3 folgen, wurde wie folgt verfahren. Die Berechnung der Matrizen *M* und *A* wurde anhand der Expressionsdaten der *Saccharomyces cerevisiae* von Gasch u. a. und Spellman u. a. durchgeführt. Die Motivmenge bestand aus 62 DNS regulatorischen Motiven und die länge einer Promotorregion wurde auf 1000 bp beschränkt. Eine Validierung der Methode wurde mittels einer Kreuzvalidierung durchgeführt, bei der die Expressionsdaten in 100 Blöcke eingeteilt wurden. Bei jedem Durchlauf wurde ein Block separiert und das Modell mit den restlichen Daten trainiert.

4.3 Ergebnisse

Ziel dieses Verfahrens ist es die Genexpression eines Gens als Funktion seiner Promotorregion zu berechnen. Für alle 5719 Gene erreicht das Modell einen durchschnittlichen Expressionskorrelationswert von 0,52 zwischen den vorhergesagten und realen Werten. Die regulatorischen Informationen, die man aus dem Modell erhalten kann, soll anhand zweier Beispiele verdeutlich werden. Die Abbildung 6 zeigt einen solchen Zusammenhang.
Die Motive *PAC* und *RRPE* besitzen bei kurzer Distanz zu dem Startcodon, hohe Werte für ihre Motivstärken. Nach diesem Modell sollten die Expressionswerte der Gene, die diese Motive in der Nähe des Startcodons haben, gut vohergesagt werden können. Dies wird durch die durchschnittliche Expressionskorrelation bestätigt. Bei einer Entfernung von 750 Basenpaaren zu *ATG* erkennt man einen erneuten Anstieg der Motivstärken, der nicht direkt durch die durchschnittliche Expressionskorrelation bestätigt werden kann. Ein *Clustering* der Gene, die das Motiv *PAC* mit einer Distanz zwischen 600 und 750 enthalten liefert hier weitere Informationen. Analysen der *Cluster* zeigen zwei unterschiedliche Expressionsklassen, die antikorreliert sind und dadurch resultiert eine durchschnittliche Expressionskorrelation von annähernd Null. Dies zeigt auch gleichzeitig, dass

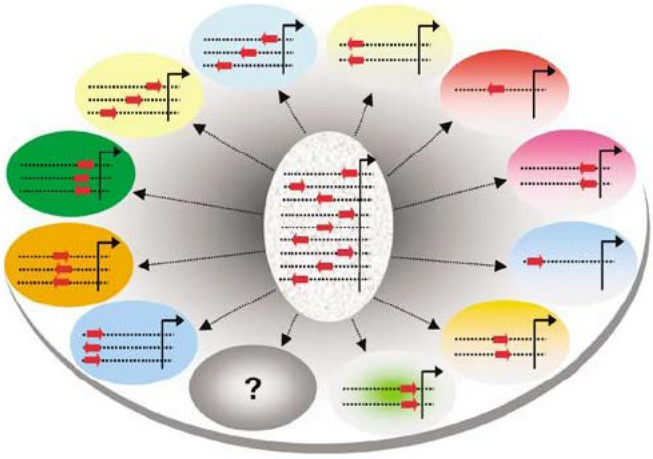

Abbildung 5: Die Abbildung zeigt das Prinzip, das den *Genensembles* zugrunde liegt. Der ovale Kreis in der Mitte stellt das *Genensemble* dar und enthält alle Gene, die die gegebene Motivmenge in ihrer Promotorregion enthalten. Die kreisförmig angeordneten ovalen Kreise repräsentieren die Instanzen des *Genensembles* und enthalten die Gene, deren Motive zusätzlich Bedingungen erfüllen. Mögliche Bedingungen neben der Geometrie sind Kombinationen von unterschiedlichen Motiven oder die Ausrichtung (Nguyen u. D'Haeseleer 2006).

die durchschnittliche Expressionskorrelation ein schwaches Maß für die Herleitung eines quantitativen Wertes darstellt (Nguyen u. D'Haeseleer 2006).
Nicht nur die Distanz zu dem Startcodon, sondern auch weitere Faktoren können aus dem Modell abgeleitet werden. So zeigt die Abbildung 7 den Zusammenhang zwischen der Motivstärke eines Motives und seiner Orientierung. Hier ist ein klarer Unterschied der Motivstärken zwischen einer $5'$ und $3'$ Orientierung zu erkennen. Die Überprüfung dieser Werte durch die durchschnittlichen Expressionskorrelationswerte bestätigt diesen Zusammenhang.
Untersuchungen von Kombinationen der Motive auf Synergismus liefern überraschende Ergebnisse. Intuitiv ist mit einem Zusammenwirken der einzelnen Motive zu rechnen, aber Daten aus der Tabelle 3 widerlegen diesen Verdacht.
Neben den genannten Funktionen, kurze Distanz zu dem Startcodon und der Orientierung, konnte das Modell noch zwei weitere Funktionen erstellen. Eine Funktion für eine mittlere Distanz (150–300 bp) und eine für große Distanzen (300–450 bp). Beide Funktionen zeigen ähnliche Resultate

	PAC $\in [ATG, -150]$ bp	*PAC* $\in [-150, -300]$ bp	*PAC* $\in [-300, -1000]$ bp
RRPE $\in [ATG, -150]$ bp	0,72	0,36	0,12
RRPE $\in [-150, -300]$ bp	0,70	0,27	0,03
RRPE $\in [-300, -1000]$ bp	0,64	0,34	-0,02

Tabelle 3: Die Tabelle zeigt mehrere Instanzen von *Genensembles*, die die Motive *PAC* und *RRPE* enthalten. Die Werte entsprechen der durchschnittlichen Expressionskorrelation von neun unterschiedlichen geometrische Konfigurationen der Motive. Die Tabelle zeigt, dass der Wert der Korrelation primär von der Position des *PAC* Motives und nur marginal von dem *RRPE* Motiv abhängt. Dies widerlegt die Vermutung eines Synergismus zwischen den beiden Motiven (Nguyen u. D'Haeseleer 2006).

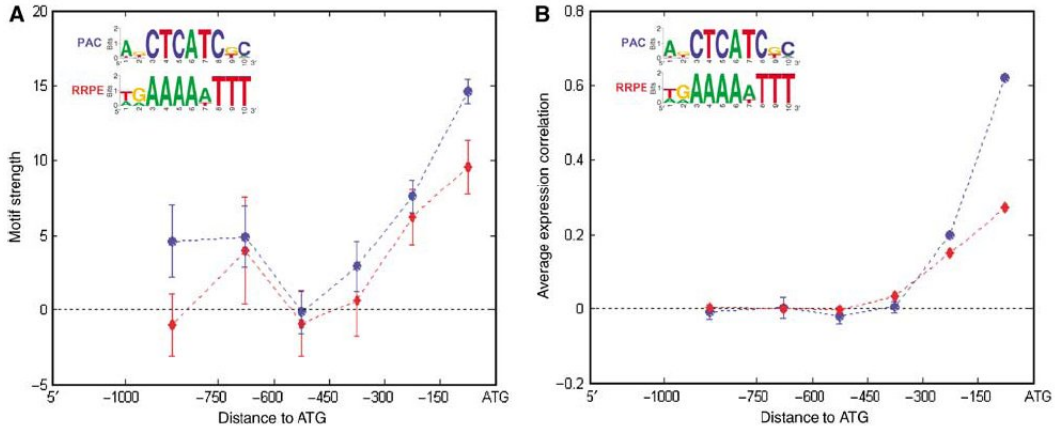

Abbildung 6: Die Abbildung zeigt den Zusammenhang zwischen der Motivstärke und den geometrischen Bedingungen für die Motive *PAC* und *RRPE*. Mit *ATG* ist die Position des Startcodons signalisiert und die Zahlen geben den Abstand zum Startcodon in Basenpaaren an. Die Abstände zum Startcodon wurde in Intervalle zu je 150 Basenpaaren zusammengefasst. Die linke Abbildung zeigt die durchschnittliche Motivstärke in Abhängigkeit von der Entfernung zu dem Startcodon. Das rechte Schaubild zeigt die durchschnittliche Expressionskorrelation als Funktion der Distanz zu dem Startcodon (Nguyen u. D'Haeseleer 2006).

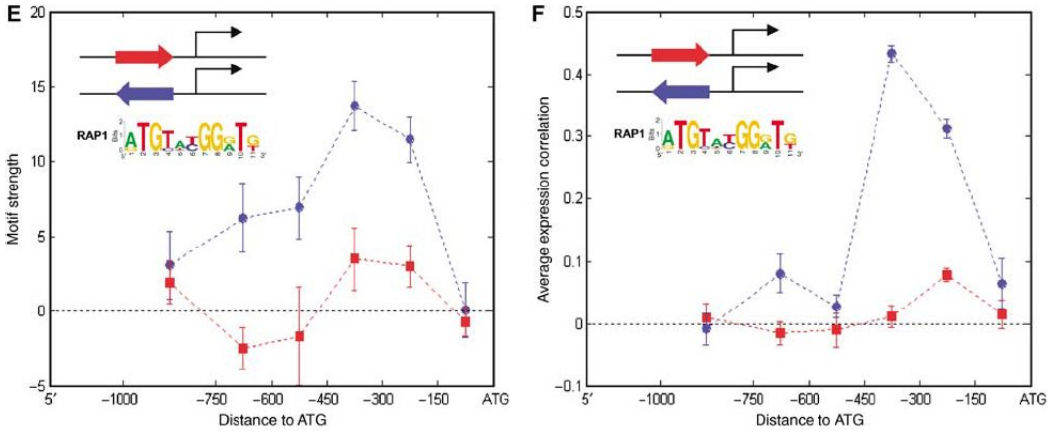

Abbildung 7: Die Abbildung zeigt den Zusammenhang zwischen der Motivstärke und den geometrischen Bedingungen für das Motiv *RAP1* mit unterschiedlicher Orientierung. Die linke Abbildung stellt die Motivstärke der beiden unterschiedlichen Orientierungen als Funktion der Distanz zu dem Startcodon dar. Auf der rechten Seite wird der Zusammenhang zwischen der durchschnittlichen Expressionskorrelation und der Entfernung zu dem Startcodon angezeigt (Nguyen u. D'Haeseleer 2006).

4.4 Fazit

Der alternative Ansatz des Modells, die Expression aus den Informationen der Motive abzuleiten, stellt eine interessante Variante dar. Ein großer Vorteil dabei ist, dass eine Reihe von möglichen Fehlerquellen ausgeschlossen werden können. Da verschiedene Transkriptionsfaktoren ihre regulierende Wirkung durch unterschiedliche biochemische Prozesse, wie Phosphorylierung, DNS-Bindung oder Bildung von Heterodimeren, entfalten können, ist eine Betrachtung der Konzentrationen der mRNS fehleranfällig. Die Ergebnisse des Modells zeigen, dass anhand der biologischen Informationen der Promotorregionen der Gene, regulatorische Funktionen abgeleitet werden können. Die Qualität der biologischen Informationen sind von elementarer Bedeutung und stellen auch einen Nachteil dar. Diese zusätzlichen Informationen müssen für den jeweils betrachteten Organismus bekannt sein.

5 Zusammenfassung

Die Inferenz korrekter regulatorischer Netzwerke stellt eine komplexe Aufgabe dar. Die unterschiedlichen Modelle und Verfahren werden mit einer Reihe von Problemen konfrontiert, die die ungenauen bzw. wenigen Resultate erklären. So beziehen alle Modelle ihre Informationen von Microarraydaten. Diese Daten Messen aber nur die Konzentrationen der mRNS und schließen daraus auf die Proteinkonzentration, was eine Korrelation zwischen den Konzentrationen der mRNS und der Proteine voraussetzt. Ein weiterer Gesichtspunkt ist die Tatsache, dass viele Modelle maschinelle Lernverfahren enthalten. Das Training dieser Verfahren sollte mit großen Datensätzen durchgeführt werden, die eine ausreichende Anzahl an Messpunkten besitzen. Diese Voraussetzung ist für die Datensätze von Spellman u. a. und Cho u. a. nicht gegeben (Spellman u. a. 1998; Cho u. a. 1998).
Dennoch haben alle vorgestellten Modelle gezeigt, dass ihre Ergebnisse biologische Relevanz haben und durchaus in der Lage sind regulatorische Zusammenhänge zu erkennen. Die Inter- und Intrazellulären Prozesse einer Zelle sind hochkomplexe biochemische Abläufe, so dass die Konstruktion aller Interaktionen nur anhand von Expressionsdaten unmöglich erscheint. Die Modelle müssen daher mehr als eine Hilfestellung angesehen werden, die Vorschläge über mögliche Interaktionen liefern, die gezielt in biologischen Versuchen getestet und verifiziert werden müssen.

Literatur

Chen u. a. 1999
CHEN, Ting ; HE, Hongyu L. ; CHURCH, George M.: Modeling Gene Expression with differential equations. In: *Pacific Symposium on Biocomputing* 4 (1999), S. 29–40

Cho u. a. 1998
CHO, R. J. ; CAMPBELL, M. J. ; WINZELER, E. A. ; STEINMETZ, L. ; CONWAY, A. ; WODICKA, L. ; WOLFSBERG, T. G. ; GABRIELIAN, A. E. ; LANDSMAN, D. ; LOCKHART, D. J. ; DAVIS, R. W.: A genome-wide transcriptional analysis of the mitotic cell cycle. In: *Molecular Cell* 2 (1998), Juli, Nr. 1, S. 65–73

Friedman u. a. 2000
FRIEDMAN, Nir ; LINIAL, Michal ; NACHMAN, Iftach ; PE'ER, Dana: Using Bayesian networks to analyze expression data. In: *RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology*. New York, NY, USA : ACM Press, 2000. – ISBN 1–58113–186–0, S. 127–135

Hughes u. a. 2000
HUGHES, J. ; ESTEP, P. ; TAVAZOIE, S. ; CHURCH, G.: Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. In: *Journal of Molecular Biology* 296 (2000), S. 1205–1214

Jacob u. Monod 1961
JACOB, F. ; MONOD, J.: Genetic regulatory mechanisms in the synthesis of proteins. In: *Journal of Molecular Biology* 3 (1961), S. 318–356

de Jong 2002
JONG, Hidde de: Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. In: *Journal of Computational Biology* 9 (2002), Januar, Nr. 1, S. 67–103

Kauffman 1969
KAUFFMAN, Stuart: Metabolic stability and epigenesis in randomly constructed genetic nets. In: *Journal of Theoretical Biology* 22 (1969), S. 437–467

Kauffman 1974
KAUFFMAN, Stuart: The Large Scale Structure and Dynamics of Gene Control Circuits. In: *Journal of Theoretical Biology* 44 (1974), S. 167–190

Kellis u. a. 2003
KELLIS, M. ; PATTERSON, N. ; ENDRIZZI, M. ; BIRREN, B. ; LANDER, E. S.: Sequencing and comparison of yeast species to identify genes and regulatory elements. In: *Nature* 423 (2003), Mai, Nr. 6937, S. 241–254. http://dx.doi.org/10.1038/nature01644. – DOI 10.1038/nature01644. – ISSN 0028–0836

Nguyen u. D'Haeseleer 2006
NGUYEN, Dat H. ; D'HAESELEER, Patrik: Deciphering principles of transcription regulation in eukaryotic genomes. In: *Molecular Systems Biology* 2 (2006), April, Nr. 1, S. msb4100054–E1–msb4100054–E10. http://dx.doi.org/10.1038/msb4100054. – DOI 10.1038/msb4100054

Segal u. a. 2003
SEGAL, E. ; SHAPIRA, M. ; REGEV, A. ; PE'ER, D. ; BOTSTEIN, D. ; KOLLER, D. ; FRIEDMAN, N.: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. In: *Nature Genetics* 34 (2003), Juni, Nr. 2, 166–176. http://dx.doi.org/10.1038/ng1165. – DOI 10.1038/ng1165. – ISSN 1061–4036

Soinov u. a. 2003
SOINOV, Lev A. ; KRESTYANINOVA, Maria A. ; BRAZMA, Alvis: Towards reconstruction of gene networks from expression data by supervised learning. In: *Genome Biology* 4 (2003), März, S. R6

Spellman u. a. 1998
SPELLMAN, Paul T. ; SHERLOCK, Gavin ; ZHANG, Michael Q. ; IYER, Vishwanath R. ; ANDERS, Kirk ; EISEN, Michael B. ; BROWN, Patrick O. ; BOTSTEIN, David ; FUTCHER, Bruce: Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. In: *Molecular Biology of the Cell* 9 (1998), Nr. 12, S. 3273–3297

Wosik 2004
WOSIK, Ewa: *Boolean Networks*, September 2004. http://cnx.org/content/m12394/1.4/.

Zhang u. Shmulevich 2002
ZHANG, W. ; SHMULEVICH, I.: Binary analysis and optimization- based normalization of gene expression data. In: *Bioinformatics* 18 (2002), Nr. 4, S. 555–565