Testdatengenerierung für die Inferenz von Genregulationsnetzen

Diplomarbeit

zur Erlangung des akademischen Grades

Diplom-Informatiker

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Fakultät für Mathematik und Informatik

eingereicht von Christian Knüpfer

(tral@minet.uni-jena.de)

Betreuer:

Prof. Dr. Clemens Beckstein

Dr. Peter Dittrich

22. März 2003

Zusammenfassung

Einer der zentralen Mechanismen in der Zelle ist die Genexpression, d.h. die Umsetzung der genetischen Information eines Gens. Das komplexe Wechselspiel zwischen Genen bei der Genexpression wird in Genregulationsnetzen dargestellt. In den letzten Jahren wurden Methoden entwickelt, die Expressionsintensitäten vieler Gene einer Zelle gleichzeitig zu messen. Mit den so gewonnenen Genexpressionsdaten ist man vielleicht in der Lage, das zugrundeliegende Genregulationsnetz aufzuklären.

Es gibt eine Reihe von Inferenzverfahren, die versuchen, aus Genexpressionsdaten ein Genregulationsnetz zu rekonstruieren. Eine Bewertung solcher Verfahren ist schwierig, da reale Genregulationsnetze im Allgemeinen unbekannt sind. Mit Modellsystemen können künstliche Genexpressionsdaten erzeugt werden. Dies ermöglicht einen direkten Vergleich zwischen dem rekonstruierten Genregulationsnetz und dem bekannten Modellsystem.

Ziel dieser Arbeit ist die Entwicklung eines Werkzeugs, welches unter Berücksichtigung bestimmter Parameter künstliche Netze erzeugt und aus diesen Genexpressionsdaten berechnet. Auf der Grundlage eines biologisch plausiblen Modells der Genexpression werden geeignete künstliche Genregulationsnetze erzeugt. Das vorgestellte Verfahren wurde in ein `Java`-Programm umgesetzt.

Mit diesem Programm werden eine Reihe von Untersuchungen vorgenommen. Es wird gezeigt, dass das Programm in der Lage ist, künstliche Systeme zu erzeugen. Diese künstlich erzeugten Systeme weisen bestimmte strukturelle Eigenschaften auf, die durch Parameter geändert werden können. Des Weiteren wird untersucht, wie sich das dynamische Verhalten der künstlichen Systeme ändert, wenn gezielte Modifikationen vorgenommen werden.

Herstellung von Genregulationsnetzen

...unter die Lupe genommen von Heike Burkhardt, einer befreundeten Künstlerin aus Jena.

Inhaltsverzeichnis

1. Einführung **7**

2. Biologischer Hintergrund **9**
2.1. Regulation der Proteinproduktion in der Zelle . . . 9
2.2. Genregulationsnetze . . . 11
2.3. Experimentell zugängliche Informationen über Genregulationsnetze . . . 15

3. Inferenz von Genregulationsnetzen **19**
3.1. Klassifikation der Inferenzverfahren . . . 20
3.1.1. Art der verwendeten experimentellen Daten . . . 20
3.1.2. Zugrundeliegender Modelltyp . . . 21
3.1.3. Art des verwendeten Verfahrens . . . 25
3.2. Bewertung von Inferenzmethoden . . . 31
3.2.1. Künstliche Genregulationsnetze . . . 32
3.2.2. Eignung der künstlich erzeugten Testsysteme . . . 33
3.2.3. Bewertungsmaße . . . 35

4. Künstliche Genexpressionsdaten **37**
4.1. Modellierung der Genexpression . . . 37
4.1.1. Komponenten und Zustand . . . 39
4.1.2. Prozesse . . . 40
4.1.3. Klassische Beschreibung . . . 42
4.1.4. Eigenschaften . . . 45
4.1.5. Dynamik . . . 47
4.2. Erzeugung künstlicher Genregulationssysteme . . . 50
4.2.1. Verfahren zur Erzeugung eines künstlichen Genregulationsnetzes . . . 51
4.2.2. Dynamisches Verhalten . . . 57
4.3. Implementierung . . . 58

5. Experimentelle Untersuchungen **63**
5.1. Erzeugung von Genregulationssystemen . . . 63
5.1.1. Vergleich einfacher und multipler GRN-Graph . . . 63
5.1.2. Regulationsmechanismus . . . 64
5.2. Verbindungsstruktur . . . 65
5.2.1. Vergleich multipler und regulatorischer Eingangsgrad . . . 65
5.2.2. Variation der Eingangsgradverteilung . . . 67
5.2.3. Selbstregulatoren . . . 68
5.3. Simulationen der Dynamik . . . 69
5.3.1. Verteilung der Genaktivitäten . . . 69
5.3.2. Aktivitätsänderungen bei einfachen Löschexperimenten . . . 71

6. Zusammenfassung **75**

A. XML-Schemata **77**
A.1. Schema für Parameterdateien . . . 77
A.2. Schema für GRN-Graphen . . . 79

Literatur **81**

1. Einführung

Einer der zentralen Mechanismen in der Zelle ist die Umsetzung der genetischen Information. Dabei werden Gene in RNS übersetzt, welche in Folge als Grundlage für die Synthese von Proteinen dienen oder andere Funktionen in der Zelle übernehmen. Der Vorgang der Übersetzung der genetischen Information eines Gens wird als Genexpression bezeichnet (Kapitel 2.1). Zellvorgänge basieren auf spezifischer Genexpression: Durch das Vorhandensein bestimmter Substanzen kann die Expression eines Gens verstärkt oder gehemmt werden. Diese Substanzen sind meist ihrerseits Genprodukte. Das komplexe Wechselspiel zwischen Genen wird in Genregulationsnetzen dargestellt (Kapitel 2.2). In den letzten Jahren wurden Methoden entwickelt, die Expressionsintensitäten vieler Gene einer Zelle gleichzeitig zu messen (Kapitel 2.3). Durch gezielte Mutationen und Zugabe von Substanzen gelingt es, diese Messungen unter verschiedenen Bedingungen zu wiederholen. Mit den so gewonnenen Genexpressionsdaten ist man vielleicht in der Lage, das zugrundeliegende Genregulationsnetz aufzuklären.

Es gibt eine Reihe von Verfahren, die versuchen, aus Genexpressionsdaten ein Genregulationsnetz zu rekonstruieren (Kapitel 3.1). Diese werden im Folgenden auch als Inferenzverfahren bezeichnet. Eine Bewertung solcher Verfahren (Kapitel 3.2) ist schwierig, da reale Genregulationsnetze im Allgemeinen unbekannt sind. Mit Modellsystemen können künstliche Genexpressionsdaten erzeugt werden. Dies ermöglicht einen direkten Vergleich zwischen dem rekonstruierten Genregulationsnetz und dem bekannten Modellsystem. Künstliche Genexpressionsdaten werden auf sehr unterschiedliche Art gewonnen. Die Bewertung einer Methode hängt in starkem Maße von den Eigenschaften der künstlichen Netze ab, wodurch ein Vergleich der Methoden untereinander sehr erschwert wird. In vielen Fällen basiert die Erzeugung künstlicher Daten auf stark vereinfachten Systemen oder auf dem selben Modell, auf dem auch das zu überprüfende Verfahren beruht. Eine mit solchen Daten durchgeführte Bewertung der Verfahren erscheint im Hinblick auf reale Messdaten fraglich.

Ziel dieser Arbeit ist die Entwicklung eines Werkzeugs, welches unter Berücksichtigung bestimmter Parameter künstliche Netze erzeugt und aus diesen Genexpressionsdaten berechnet. Auf der Grundlage eines biologisch plausiblen Modells der Genexpression (Kapitel 4.1) werden geeignete künstliche Genregulationsnetze erzeugt (Kapitel 4.2). Das dynamische Verhalten dieser künstlichen Genregulationsnetze wird durch Differenzialgleichungssysteme beschrieben, mit deren Hilfe künstliche Genexpressionsdaten gewonnen werden können. Damit ist eine vergleichende Bewertung von Inferenzverfahren möglich. Durch geeignete Wahl der Parameter können die Auswirkungen bestimmter Annahmen auf die Ergebnisse eines Inferenzverfahrens untersucht werden. Das vorgestellte Verfahren wurde in ein `Java`-Programm umgesetzt (Kapitel 4.3).

Mit diesem Programm wird eine Reihe von Untersuchungen vorgenommen. Es wird gezeigt, dass das Programm in der Lage ist, entsprechende künstliche Systeme zu erzeugen (Kapitel 5.1). Diese künstlich erzeugten Systeme weisen bestimmte strukturelle Eigenschaften auf, die durch Parameter geändert werden können (Kapitel 5.2). Des weiteren wird untersucht, wie sich das dynamische Verhalten der künstlichen Systeme ändert, wenn gezielte Modifikationen vorgenommen werden (Kapitel 5.3). Die Simulation der Dynamik eines künstlichen Systems entspricht der Messung von Genexpressionsdaten unter verschiedenen Bedingungen.

2. Biologischer Hintergrund

Dieses Kapitel soll kurz das zum Verständnis dieser Arbeit notwendige biologische Wissen umreißen und den zentralen Begriff des Genregulationsnetzes einführen. Es wird eine Übersicht über den Kenntnisstand grundlegender Eigenschaften dieser Netze gegeben. Diese Beschreibung der biologischen Realität dient als Basis für die zu generierenden künstlichen Regulationsnetze (Kapitel 4). Außerdem werden die experimentellen Methoden vorgestellt, die die Daten für die Inferenz von Genregulationsnetzen liefern.

Die meisten der hier verwendeten biologischen Begriffe können z.B. in Sauermost (1994) oder Sauermost (1991-1992) nachgeschlagen werden.

2.1. Regulation der Proteinproduktion in der Zelle

In diesem Abschnitt erfolgt eine komprimierte Darstellung der Vorgänge und Regulationsmechanismen bei der Übersetzung der genetischen Information in der Zelle. Die dabei stattfindenden regulierenden Wechselwirkungen zwischen Genen werden in Genregulationsnetzen dargestellt, womit sich der Abschnitt 2.2 auseinandersetzt. Detailliertere Informationen zur Regulation der Genexpression sind in den meisten Lehrbüchern der Genetik zu finden, so z.B. Hagemann (1991). Einen gut verständlichen Einblick in das derzeitig gesicherte Wissen gibt Munk (2001) mit einem jeweils eigenen Kapitel über die Vorgänge bei der Transkription (Kapitel 4), bei der Translation (Kapitel 5) und über die Regulation der Genexpression (Kapitel 10).

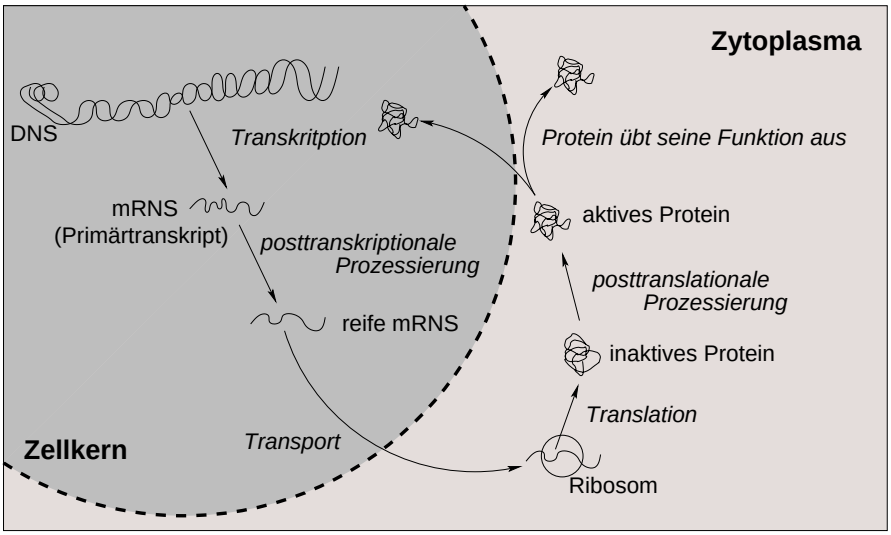

Abb. 2.1.: Schematische Darstellung der Genexpression eines eukaryotischen Strukturgens nach Lehninger et al. (1994).

Bei Prokaryoten und Eukaryoten ist die genetische Information in Form von *DNS* gespeichert, bei einigen Viren zum Teil auch als RNS. Die Gesamtheit der genetischen Information einer Zelle wird als *Genom* bezeichnet. Ein *Gen* ist eine Teilsequenz der DNS, die entweder die Information für die Synthese eines Proteins enthält (Strukturgen) oder ein RNS-Molekül kodiert, das eine bestimmte Funktion in der Zelle ausübt. Der Vorgang der Realisierung der Funktion eines Gens wird als *Genexpression* bezeichnet.

Die Genexpression

erfolgt in mehreren Teilschritten (Abbildung 2.1). Zuerst wird die entsprechende Teilsequenz der DNS in RNS übertragen. Dieser Vorgang wird als *Transkription* bezeichnet. Bei Strukturgenen schließt sich die Übersetzung der RNS-Sequenz in eine Folge von Aminosäuren an, das entsprechende Protein wird aufgebaut (*Translation*). RNS, die als Grundlage für die Proteinsynthese dient, wird als Boten-RNS bzw. mRNS (*Messenger*-RNS) bezeichnet. Bei Prokaryoten erfolgt die Proteinsynthese direkt am Ort der Transkription und beginnt noch bevor diese abgeschlossen ist. Demgegenüber wird bei Eukaryoten die fertig transkribierte RNS durch bestimmte Modifikationen in ihre funktionell aktive Form überführt und aus dem Zellkern in das Zytoplasma transportiert, wo die Proteinsynthese im Ribosom stattfindet.

Regulation der Genexpression

Trotz identischer genetischer Information weisen die Zellen eines Organismus in Abhängigkeit von inneren oder äußeren Reizen, vom Entwicklungsstadium, von der Phase im Zellzyklus und (bei mehrzelligen Organismen) vom Zelltyp völlig unterschiedliches biochemisches Verhalten auf. Dabei spielt die Kontrolle der Proteinsynthese (und damit der Genexpression) die zentrale Rolle bei der Steuerung der biochemischen Prozesse (Munk, 2001, S. 10-1). Proteine, die der momentanen Situation der Zelle angemessen sind, werden synthetisiert, die Synthese nicht benötigter Proteine wird zur Vermeidung von Energieverschwendung unterdrückt. Das bedeutet, dass in einer Zelle zu einem bestimmten Zeitpunkt nur ein gewisser Teil der Gene des Genoms exprimiert wird.

Die Kontrolle der Proteinsynthese kann prinzipiell auf allen Stufen der Genexpression erfolgen. Vor der Transkription kann Regulation durch Sequenzänderung, Verlust, Vervielfältigung und Zustandsmodifikation der DNS bzw. Teilen davon bewirkt werden. Gerade in Eukaryoten gibt es eine Reihe von posttranskriptionalen Regulationsmechanismen. Eine der wichtigsten Kontrollebenen stellt jedoch die Transkription dar. Da die in dieser Arbeit behandelten Inferenzverfahren für Genregulationsnetze vor allem auf dieser Ebene operieren (siehe Kapitel 3), wird sich hier auf eine kurze Abhandlung der Regulation der Transkription beschränkt. Für Details sei wiederum auf Munk (2001) verwiesen.

Regulation der Transkription

Die Transkription kann in drei Phasen unterteilt werden: Start der Transkription (*Initiation*), Synthese der RNS (*Elongation*) und Beendigung der Transkription (*Termination*). Bei der Initiation bindet ein Molekülkomplex (RNS-Polymerase), der für die Synthese der RNS zuständig ist, an einen bestimmten Bereich des Gens, den *Promotor*. Dies ist die entscheidende Kontrollstelle bei der Transkription: Ob und wie oft die RNS-Polymerase an den Promoter bindet und somit die Elongation startet.

Die Häufigkeit der Initiation wird durch DNS-Abschnitte, die als *cis-Elemente* bezeichnet werden, reguliert. An diese Stellen binden spezifische Proteine und verstärken (*Enhancer*) oder vermindern (*Silencer*) die Transkriptionshäufigkeit. Diese Proteine werden als *trans-Faktoren* bezeichnet. In der Literatur werden die trans-Faktoren häufig mit den *Transkriptionsfaktoren* gleichgesetzt. Nach Munk (2001) stellen die Transkriptionsfaktoren jedoch nur eine spezielle Art von trans-Faktoren dar.

Die Proteine, die an die cis-Faktoren binden, können ihrerseits Produkte von exprimierten Genen sein. Die Expression eines Gens hängt also von der Expression anderer Gene ab und beeinflusst die Expression anderer Gene. Diese gegenseitigen Wechselwirkungen werden in Form eines Netzwerkes veranschaulicht. Dieses wird als *Genregulationsnetz* bezeichnet.

2.2. Genregulationsnetze

Ein Genregulationsnetz ist die abstrakte Darstellung der regulierenden Wechselwirkungen zwischen den Genen eines Genoms. Die Gene werden durch Knoten repräsentiert, der Einfluss der Expression eines Gens auf die Expression eines anderen Gens durch eine gerichtete Kante zwischen den zugehörigen Knoten. Die Kanten können zusätzlich mit einer Kennzeichnung versehen werden, die die Art (hemmend oder verstärkend) oder die Stärke des Einflusses (numerischer Wert) angibt. Das Genregulationsnetz kann darüber hinaus auch Knoten für externe Zellsignale enthalten, die nur ausgehende Kanten besitzen, also Quellknoten sind. Will man posttranskriptionale Regulationsprozesse oder den Einfluss von Protein-Protein-Wechselwirkungen auf die Genexpression explizit darstellen, können zusätzlich Knoten für Proteine eingeführt werden. D'haeseleer (2000) schlägt ein solches zweischichtiges Netz vor.

Eigenschaften von Genregulationsnetzen

Ziel dieser Arbeit ist es, für die Bewertung von Inferenzverfahren künstliche Genregulationsnetze zu erzeugen. Diese künstlichen Netze sollen in bestimmten Merkmalen realen Genregulationsnetzen möglichst ähnlich sein. Als Grundlage dafür dienen die im Folgenden zusammengetragenen Eigenschaften von Genregulationsnetzen.

Durch Analyse der in Datenbanken und der Literatur verfügbaren Informationen über die Regulation der Genexpression in speziellen Organismen können grundsätzliche Eigenschaften von Genregulationsnetzen abgeleitet werden. So untersuchten Thieffry et al. (1998) die bekannten Teile des transkriptionalen Genregulationsnetzes von *E. coli* und fanden bestimmte topologische Merkmale.

Eine andere Informationsquelle stellt die DNS-Sequenz dar. Da die Kontrolle der Transkription durch Bindung von Proteinen an cis-Elemente erfolgt (siehe 2.1), können durch Suche nach cis-Elementen auf der DNS-Sequenz und Identifikation der Gene, die für die zugehörigen Proteine kodieren, Aussagen über Teile des Genregulationsnetzes gewonnen werden. Arnone und Davidson (1997) analysierten die cis-Elemente von Genen unterschiedlicher Organismen und leiteten daraus strukturelle Eigenschaften von Genregulationsnetzen ab.

Jeong et al. (2000) untersuchten das Datenbankwissen über die metabolischen Netzwerke von 43 verschiedenen Organismen im Hinblick auf ihre globalen Organisationsstrukturen. Dabei fanden sie Ähnlichkeiten zu anderen komplexen dynamischen Netzwerken, wie dem Internet und sozialen Strukturen. Alle diese Systeme weisen bestimmte gemeinsame Merkmale auf, zum Beispiel Robustheit und Fehlertoleranz, die durch ähnliche topologische Prinzipien hervorgerufen werden. Die Verteilung der Konnektivität der Knoten in solchen Netzen folgt Potenzfunktionen: Die Wahrscheinlichkeit $P(k)$, dass ein Knoten eine Konnektivität von k hat, ist gegeben durch $P(k) = \alpha k^{-\gamma}$ mit Konstanten $0 < \alpha \leq 1$ und $\gamma > 0$. Die Form der grafischen Darstellung ist invariant gegenüber α, weshalb vernetzte Systeme, deren Konnektivität einer Potenzgesetzverteilung folgen, als *scale-free Networks* bezeichnet werden (Barabási und Albert, 1999). Da Genregulationsnetze ein analoges Verhalten zeigen, liegt die Vermutung nahe, dass ähnliche Organisationsstrukturen zugrunde liegen. Das wachsende Wissen über Genregulationsnetze sollte eine Überprüfung dieser Vermutung ermöglichen.

Auf der Grundlage von Simulationen mit zufällig erzeugten booleschen Netzen (*Random Boolean Network*) stellte Kauffman (1993) theoretische Überlegungen über notwendige Eigenschaften von Genregulationsnetzen an. Unter welchen Randbedingungen erzeugen die

künstlichen Netze ein Verhalten, das in bestimmten Merkmalen den biologischen Systemen ähnlich ist? Ob diese Eigenschaften auch biologisch notwendig sind, bleibt ein offenes Problem. Somogyi und Sniegoski (1996) setzten diese theoretischen Untersuchungen mit Hilfe von booleschen Netzen fort.

Die Entwicklung neuer Messmethoden und Verfahren zur Untersuchung der Genregulation in den letzten Jahren sorgt für ein sich stückweise aufbauendes Bild der Genregulationsnetze. Dieses ist jedoch noch weit davon entfernt, vollständig zu sein. Trotzdem seien im Folgenden einige bekannte Eigenschaften von Genregulationsnetzen genannt. Diese sind mit entsprechender Vorsicht zu bewerten.

Größe: Die geschätzte Anzahl von Genen in den Genregulationsnetzen der meisten Organismen liegen zwischen 3237 bei *E. coli* (Bakterium), 6034 bei *S. serevisiae* (Hefe) und in etwa 30000 bei Säugetieren (`http://www.ornl.gov/hgmis/faq/compgen.html`).

Will man außertranskriptionale Regulation explizit darstellen, erhöht sich die Größe des Genregulationsnetzes etwa um den Faktor 10 (Szallasi, 1999).

Eingangsgrad: Die meisten Gene werden nur von einer relativ kleinen Anzahl von Proteinen reguliert. In *E. coli* liegt nach Thieffry et al. (1998) der mittlere (transkriptionale) Eingangsgrad bei zwei, die maximale Anzahl von sechs regulierenden Proteinen wird nur bei einem einzigen Gen erreicht. Die von Arnone und Davidson (1997) untersuchten Gene wiesen vier bis acht (im Mittel fünf) unterschiedliche cis-Elemente auf, somit kommt also eine Regulierung durch etwa fünf Proteine in Frage.

Vernachlässigt man nichttranskriptionale Regulation, kann man daraus auf einen Eingangsgrad im Genregulationsnetz in dieser Größenordnung schließen. Die Aktivität eines Proteins wird jedoch von einer Reihe anderer Prozesse gesteuert, die ebenfalls durch Genprodukte beeinflusst werden können. Dies kann zu weiteren Kanten im Genregulationsnetz und somit zu wesentlich höheren Eingangsgraden führen. Repräsentiert man die nichttranskriptionalen Regulationsereignisse durch eigene Knoten, kommt es nicht zu einer Erhöhung des Eingangsgrads. Die Größe des Genregulationsnetzes erhöht sich dadurch jedoch erheblich (siehe oben und Szallasi, 1999).

Thieffry und Thomas (1998) geben ein Indiz für einen kleinen Eingangsgrad bei *E. coli* an: die Mechanismen bei der Initiation der Transkription erfordern eine relative Nähe von Promotor und cis-Elementen. Somit kann jeder Promotor nur von einer geringen Anzahl von cis-Elementen gesteuert werden. Diese Aussage trifft allerdings auf Eukaryoten nicht zu.

Kauffman (1993) zeigt, dass zufällig erzeugte boolesche Netzwerke mit Eingangsgrad zwei das am biologisch plausibelste Verhalten zeigen und dass, je höher der Eingangsgrad gewählt wird, die Merkmale von biologischen Genregulationsnetzen immer mehr verschwinden.

Man trifft auch auf starke Kritik an der Annahme eines beschränkten Eingangsgrads für alle Gene: Weaver et al. (1999, S. 2) führen an, dass unser Wissen über Genregulationsnetze zu begrenzt ist, um solche Aussagen treffen zu können. Für einige Gene sei ein sehr großer Eingangsgrad bekannt, während man über andere nicht wüsste, ob sie mehr als ein paar regulatorische Eingänge hätten. Wahrscheinlich seien die Eingangsgrade der Gene zu stark verteilt, um eine solche Beschränkung rechtfertigen zu können.

Der Begriff „Eingangsgrad“ ist hier unscharf definiert und wird auch in der Literatur nicht einheitlich benutzt. So definiert die Arbeitsgruppe Akutsu et al. den Eingangsgrad als Anzahl von Genen, die einen direkten Einfluss auf die Expression eines Gens

ausüben (Akutsu et al., 1998a, S. 2; Akutsu et al., 1998b, S. 154). Diese Definition ist nur sinnvoll, wenn genaue Kenntnisse über alle Vorgänge in einem Genregulationsnetz vorliegen, also zwischen direktem und indirektem Einfluss unterschieden werden kann.

Ausgangsgrad: Auch die Anzahl der Gene, die von einem bestimmten Gen reguliert werden, ist relativ klein. Thieffry et al. (1998) fanden zwar eine wesentlich breitere Verteilung des Ausgangsgrads im Genregulationsnetz von *E. coli*, dieser liegt mit drei im Mittel jedoch nur knapp über dem mittleren Eingangsgrad. Nur wenige Gene, die die Rolle von Hauptregulatoren spielen, haben einen deutlich höheren Ausgangsgrad (Maximum 133).

Konnektivität: Die Untersuchungen von Jeong et al. (2000) zeigen, dass sowohl die Eingangs-, wie auch die Ausgangsgrade in metabolischen Netzen Potenzgesetzverteilungen folgen, die für *scale-free Networks* typisch sind (siehe oben). Dieses Ergebnis lässt sich möglicherweise auf Genregulationsnetze übertragen (Jeong et al., 2000), da beiden Arten von Netzwerken ähnliche Entwicklungsmechanismen zugrunde liegen (Barabási und Albert, 1999). Bei diesen Untersuchungen werden Knoten, die einen Eingangs- bzw. Ausgangsgrad von Null besitzen, jeweils nicht in die Häufigkeitsanalyse aufgenommen.

Savageau (1998, S. 10) führt Gründe für einen geringen Grad an Konnektivität an: Molekulare Analysen hätten gezeigt, dass ein Gen nur von einer geringen Anzahl von Regulatorsubstanzen beeinflusst wird. Zu dem selben Ergebnis kämen auch Untersuchungen von Regulatorsubstanzen und den zugehörigen cis-Elementen auf der Grundlage von Sequenzhomologien. Als drittes Argument führt er die theoretischen Untersuchungen an zufällig erzeugten Netzen von Kauffman (1993) an, nach denen schwach verbundene Netze das biologisch plausibelste Verhalten aufwiesen.

Hubs: Eine Folge der Potenzgesetzverteilung der Konnektivität ist die Existenz von wenigen (geringe Wahrscheinlichkeit) Knoten mit hohem Eingangs- oder Ausgangsgrad, die eine dominierende Rolle im Netz übernehmen und als *Hubs* bezeichnet werden (Jeong et al., 2000). So fanden Thieffry et al. (1998) wenige Gene mit hohem Ausgangsgrad, die eine wichtige regulative Rolle im Genregulationsnetz von *E. coli* spielen.

Rückkopplungen: Rückkopplungen stellen einen wichtigen Mechanismus dar, um Differenzierung in verschiedene Zelltypen und stabile Zellzustände (*Homöostase*) zu ermöglichen. Thieffry und Thomas (1998) fassen die Untersuchungen zur Rolle von Rückkopplungen in dynamischen Systemen zusammen: „Positive“ Rückkopplungen (gerade Anzahl negativer Regulierungen im Rückkopplungskreis) führen zu multiplen stationären Zuständen, was eine biologische Entsprechung in der Zelldifferenzierung findet. Dagegen neigen „negative“ Rückkopplungen zu periodischem Verhalten, was der biologischen Homöostase entspricht. Die überwiegende Mehrheit der Rückkopplungen im Genregulationsnetz von *E. coli* ist einstufig, also Selbstregulationen von Genen (Thieffry et al., 1998) (siehe auch nächster Punkt).

Selbsregulation: Viele Gene regulieren ihre eigene Expression. Fast die Hälfte der Gene von *E. coli* ist nach Thieffry et al. (1998) selbstregulierend und die meisten Selbstregulationen sind Autoinhibitionen.

Modularität: Ausgehend von den Untersuchungen zum Genregulationsnetz von *E. coli* (Thieffry et al., 1998) und Analysen des dynamischen Verhaltens von booleschen Netzen schließen Thieffry und Thomas (1998), dass Genregulationsnetze aus wenigen kleinen und untereinander schwach vernetzten Teilnetzen (Modulen) aufgebaut sind. Innerhalb der Module herrscht eine stärkere Vernetzung und ein Modul realisiert eine bestimmte biologische Funktion („unabhängige Regulationswege“: Thieffry und Thomas, 1998, S. 10). Deutlich wird diese modulare Struktur, wenn man einzelne Gene eines Organismus genetisch deaktiviert. Meist wird dadurch nur die Expression einer geringen Anzahl anderer Gene beeinflusst (Thieffry et al., 1998).

Redundanz: Nach Yeung et al. (2002) gibt es Hinweise, dass Genregulationsnetze über multiple Pfade zur Realisierung derselben biologischen Funktion verfügen, die es einer Zelle z.B. beim Ausfall eines Gens ermöglicht, einen alternativen Weg zur Synthese einer benötigten Substanz zu beschreiten. Somit sei die für die Clusteranalyse vorausgesetzte eindeutige Zuordnung zwischen Expression und Funktion eines Gens fraglich.

Struktur nicht hierarchisch: Arnone und Davidson (1997) fanden Hinweise, dass ein Genregulationsnetz nicht streng hierarchisch angeordnet sein kann, d.h. eine Einteilung der Gene in Schichten ist nicht möglich. Dabei wird unter einer Schicht eine Menge von Genen verstanden, die nur Verbindungen aus der darüberliegenden Schicht und zur darunterliegenden Schicht besitzen. Da zwischen zwei Genen unterschiedliche Pfade existieren können, die miteinander konkurrieren und verschiedene Anzahlen von Zwischenstufen aufweisen, ist eine solche Schichtung nicht möglich. Diese nicht hierarchische Struktur führen Yeung et al. (2002) als Problem bei der hierarchischen Clusterung von Genexpressionsdaten an.

Mittlere Pfadlänge: Die Untersuchung der metabolischen Netze von Jeong et al. (2000) zeigte, dass die mittlere kürzeste Pfadlänge zwischen je zwei Substanzen im Netz (Diameter) unabhängig von der Größe des Netzes ist. Dieses Ergebnis sei unerwartet: Bei allen betrachteten nicht-biologischen komplexen Netzwerken sei die mittlere Konnektivität eines Knotens unabhängig von der Netzgröße. Dies hat zur Folge, dass der Diameter logarithmisch mit der Netzgröße ansteigt. Es hat den Anschein, dass mit zunehmender Komplexität des Organismus (und damit steigender Anzahl Gene) die mittlere Konnektivität der Knoten zunimmt und somit der Diameter konstant bleibt (Jeong et al., 2000).

Stochastizität: Bei der Regulation der Genexpression spielen zum Teil Substanzen eine Rolle, von denen nur wenige Moleküle pro Zelle vorhanden sind. In solchen Fällen stellt eine deterministische Beschreibung der Prozesse keine gute Näherung mehr dar, da der Zustand einer Zelle in starkem Maße von zufälligen Ereignissen abhängt (D'haeseleer, 2000, S. 43). Nach Szallasi (1999) gibt es für die stochastische Natur sowohl theoretische als auch experimentelle Belege.

Kanalisation: Von Kauffman (1993) wurden Simulationen mit zufällig erzeugten booleschen Netzen durchgeführt. Falls den Knoten ein bestimmter Typ boolescher Funktionen zugeordnet wird, zeigen die Netze ein biologisch plausibles Verhalten. Dieser Funktionstyp (*canalyzing Function*) hat die Eigenschaft, dass eine der Eingangsvariablen eine Belegung besitzt, die einen bestimmten Wert der Funktion (unabhängig von den anderen Eingängen) erzwingt. Funktionen wie das exklusive Oder (`XOR`) fallen nicht in diese Kategorie. Biologisch lässt sich dies so interpretieren: Besitzt ein Gen zwei cis-Elemente, an die zwei verschiedene Aktivatoren binden können, so erscheint es unwahrscheinlich, dass das Binden von jeweils einem Aktivator die Expression des Gens verstärkt, während ein gleichzeitiges Binden beider Aktivatoren die Expression verhindert.

D'haeseleer et al. (1999) modellierten das Genregulationsnetz der Entwicklung des zentralen Nervensystems der Ratte mit linearen Differenzialgleichungssystemen, wobei die Expressionsrate eines Gens als gewichtete Summe der Konzentrationen der beeinflussenden Gene ausgedrückt wird. Sie fanden dabei Eigenschaften, die sie als biologisch plausibel bezeichnen:

- Die Konnektionsmatrix ist dünn besetzt (geringe Konnektivität).
- Die Summe der Eingangsgewichte der meisten Gene ist annähernd Null, d.h. die meisten Gene können in ihrer Expression von anderen Genen sowohl gehemmt als auch verstärkt werden.
- Die meisten Gene können sowohl hemmende als auch aktivierende Einflüsse ausüben.
- Einige wenige Gene üben bevorzugt einen hemmenden bzw. aktivierenden Einfluss aus.
- Die Grundaktivität der Gene bei Abwesenheit von geeigneten Regulatoren ist gering.

2.3. Experimentell zugängliche Informationen über Genregulationsnetze

Um die oben beschriebenen Wechselwirkungen in Genregulationsnetzen aufzudecken, ist es notwendig, bestimmte Zustandsgrößen der Zelle zu messen. Hier erfolgt eine Darstellung der dafür zur Verfügung stehenden experimentellen Methoden.

Der erste Schritt bei der Untersuchung des Genregulationsnetzes eines Organismus besteht in der Bestimmung der Nukleinsäuresequenz des Genoms. Für die *Sequenzierung* existieren grundsätzlich zwei Verfahren: Maxam-Gilbert-Technik und Sanger Verfahren, die in den letzten Jahren stark weiterentwickelt und automatisiert wurden (Munk, 2001). Durch Analyse der so gewonnenen Daten können die im Organismus vorliegenden Gene bestimmt werden. Beide Schritte zusammen bezeichnet man als *Sequenzanalyse*. Zahlreiche Genomprojekte, deren Aufgabe in der Sequenzanalyse verschiedener Organismen liegt, konnten in jüngster Vergangenheit zum Abschluss gebracht werden bzw. sind kurz vor ihrer Vollendung (Bernal et al., 2001).

Man kann also davon ausgehen, dass die Knoten (Gene) eines Genregulationsnetzes mit Hilfe der Sequenzanalyse bekannt sind. Die Herausforderung liegt somit in der Aufklärung der Regulationsvorgänge bei der Genexpression, also im Auffinden der Kanten des Genregulationsnetzes.

Es existieren eine Reihe von biochemischen Verfahren, durch die kleine Ausschnitte des Genregulationsnetzes im Detail untersucht werden können. Auf diese „im kleinen Maßstab“ wirkenden Verfahren soll hier nicht eingegangen werden.

Demgegenüber stehen Verfahren, die im „großen Maßstab“ (*Large-Scale*) Daten über das Genregulationsnetz liefern. So wurden Mitte der 90er Jahre Methoden entwickelt, die es erlauben, die Konzentration von mehreren tausend mRNS gleichzeitig zu messen. In diesem Zusammenhang sprechen D'haeseleer et al. (1999) von einer Verlagerung des Schwerpunktes der Forschung von der Sequenzanalyse ganzer Genome (*structural Genomics*) hin zur Aufklärung der Funktionsweise der Realisierung genetischer Information (*functional Genomics*). Eines der wichtigsten Mittel der *functional Genomics* sei die Messung der Genexpression im großen Maßstab.

Durch detaillierte Analyse der vorliegenden DNS-Sequenz können weitere Information gewonnen werden, wie z.B. Informationen über das kodierte Protein. Arnone und Davidson (1997) analysierten die Struktur von cis-Elementen und schlossen daraus auf einen Teil der transkriptionalen Regulationen.

Da die Regulation der Genexpression oft über Proteine oder kleinere Moleküle (Metabolite) vermittelt wird, stellt die Messung von Protein- und Metabolitkonzentrationen ebenfalls eine wichtige Informationsquelle für die Untersuchung von Genregulationsnetzen dar. Zur Zeit gibt es Bestrebungen solche Messungen im großen Maßstab durchzuführen, allerdings ist die Technik hierfür noch nicht so weit fortgeschritten wie bei der Messung von mRNS-Konzentrationen.

Simultane Messung der mRNS-Konzentrationen vieler Gene

Die momentan am häufigsten verwendeten Daten zur Inferenz von Genregulationsnetzen stammen aus der parallelen Messung der Konzentration sehr vieler (bis zu mehreren tausend) mRNS. Der Grund dafür liegt in der Verfügbarkeit: die in den 90er Jahren entwickelten *Microarray*-Techniken gestatten die Messung von mRNS-Konzentrationen in großem Maßstab mit vergleichsweise geringem Aufwand, während zum Beispiel die Messung vieler Protein-Konzentrationen mit erheblichen Schwierigkeiten verbunden ist (D'haeseleer, 2000). Nach Lander (1996) beschreiben die mRNS-Konzentrationen aller Gene ziemlich gut - möglicherweise eindeutig - den Zustand einer Zelle, sind also ein geeignetes Mittel zur Aufklärung der Regulation der Genexpression.

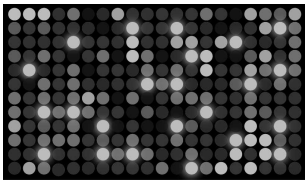

Abb. 2.2.: Beispiel eines Microarrays, Quelle: U.S. Department of Energy Genomes to Life Program, `http://doegenomestolife.org`.

Grundsätzlich existieren zur Zeit zwei Verfahren zur simultanen Messung von mRNS-Konzentrationen: Die „DNA-Chips“, die von der Firma Affymetrix entwickelt wurden und die „cDNA-Microarrays“, entwickelt an der Stanford-Universität. Beide werden häufig jedoch einfach als „Microarrays“ bezeichnet.

Beiden Verfahren gemeinsam ist das Messprinzip: Auf einer Oberfläche sind einfachsträngige DNS-Sequenzen aufgebracht, an denen sich aus der Probe entsprechende komplementäre RNS anlagern und so stabile mRNS/DNS-Hybride bilden können. Da die mRNS vorher mit fluoreszierender Farbe markiert wird, kann die Hybridisation mit Hilfe der Fluoreszenz gemessen werden. Die Stärke der Fluoreszenz ist somit proportional zur Konzentration der entsprechenden mRNS (D'haeseleer, 2000). In der Praxis wird die mRNS zuvor meist in stabilere cDNS (komplementäre DNS) zurückübersetzt („reverse-transcripted“). Auf dem Microarray befindet sich zu dieser cDNS komplementäre DNS (D'haeseleer, 2000). Da auf einem Microarray viele verschiedene DNS-Sequenzen aufgebracht sind, ist die simultane Messung der Konzentrationen von bis zu mehreren tausend mRNS möglich. Abbildung 2.2 illustriert eine solche Messung: jeder Farbpunkt entspricht einer bestimmten DNS-Sequenz, die Farbintensität ist ein Maß für die Konzentration der zugehörigen mRNS in der Probe.

Es existieren verschiedene Wege, wie ein Microarray hergestellt werden kann. Der Hauptunterschied zwischen den beiden genannten Verfahren liegt gerade in der Art und Weise, wie die DNS-Stränge erzeugt werden: Bei den cDNA-Microarrays wird eine Probe von DNS-Strängen biochemisch vervielfältigt und anschließend an der entsprechenden Position auf das Array gebracht. Bei den DNA-Chips wird die DNS-Sequenz direkt auf der Oberfläche aus den Nukleinsäuren synthetisiert. Dafür muss die Sequenz vorher bekannt sein (Dutilh und Hogeweg, 1999).

Ein Problem bei der Messung mittels Microarrays stellt die Aufbereitung der Probe dar: Optimal wäre die Gewinnung aus einer einzelnen Zelle, was aber technische Probleme mit sich bringt. Viel häufiger wird daher mit ganzen Populationen von Zellen eines Typs bzw. einer Gewebeart gearbeitet. Damit erhält man eine Durchschnittsmessung über alle Zellen, womit die Untersuchung von intrazellulärer Signalverarbeitung stark erschwert wird (Smolen et al., 2000). Außerdem kann es zu Fehlinterpretationen kommen, wenn z.B. Gene für co-exprimiert gehalten werden, die sich eigentlich gegenseitig in ihrer Expression ausschließen (Dutilh und Hogeweg, 1999). Da sich die Zellen einer Population auch in unterschiedlichen Phasen im Zellzyklus befinden, wird für Gene, die eine zyklische Expression aufweisen nur eine mittlere Konzentration gemessen, die Information über das Zeitverhalten der Expression geht verloren. Weiterhin führt Szallasi (1999) an, dass durch Messung von Zellpopulationen stochastische Effekte bei der Regulation der Genexpression verdeckt werden können. Als möglichen Lösungsweg gibt er stochastische Simulationen an.

Neben den Microarray-Methoden existieren weitere Verfahren zur Messung der mRNS-Konzentration, wie z.B. SAGE und RT-PCR (siehe D'haeseleer, 2000). Diese zeichnen sich durch eine höhere Genauigkeit aus, sind aber sequentiell arbeitende Methoden. Durch Automatisierungstechniken gelingt es jedoch seit einigen Jahren, mit diesen Verfahren in kurzer Zeit ebenfalls die Konzentrationen vieler mRNS zu messen.

Von einer Zellkultur können in bestimmten Zeitabständen Proben genommen und einer Genexpressionsmessung unterzogen werden. Somit erhält man Zeitserien der Genexpression, was besonders zur Verfolgung von Entwicklungsprozessen geeignet ist. Ein anderes Vorgehen stellt die Messung während eines stabilen Zustands der Zellkultur dar (*Steady-State*). Dies kann unter verschiedenen Bedingungen, wie der gentechnischen Deaktivierung einzelner Gene oder der Zugabe von bestimmten Substanzen, durchgeführt werden. Der Informationsgehalt der gewonnenen Daten kann mit solchen „Störexperimenten“ weiter erhöht werden.

3. Inferenz von Genregulationsnetzen

Im vorangegangenen Kapitel wurden biologische Genregulationsnetze vorgestellt und gezeigt, welche experimentellen Daten über diese gewonnen werden können. Hier werden Verfahren behandelt, die von solchen Daten auf das zugrundeliegende Genregulationsnetz schließen. Dabei liegt der Schwerpunkt auf Verfahren, deren Hauptinformationsquelle Microarray-Daten darstellen und deren Ziel in der Aufdeckung von regulatorischen Einflüssen zwischen Genen liegt. Diese werden im Folgenden als **Inferenzmethoden** bezeichnet.

Von der Betrachtung ausgenommen sind zum einen Verfahren, die überwiegend andere Daten benutzen, etwa biologisches Wissen aus Datenbanken (de Jong et al., 2001), Sequenzinformationen (Tavazoie et al., 1999; Yuh et al., 1998) oder beide Arten kombiniert (Davidson et al., 2002). Zum anderen werden Verfahren, die hauptsächlich andere Ziele als die Aufdeckung des zugrundeliegenden Genregulationsnetzes verfolgen, in dieser Untersuchung nicht betrachtet. Dies sind z.B. Clustermethoden, d.h. die Einteilung von Genen in funktionelle Gruppen auf der Grundlage von Ähnlichkeiten in ihren Expressionsprofilen (für eine kurze Übersicht über Clustermethoden siehe D'haeseleer et al., 2000) und Untersuchungen von grundsätzlichen Eigenschaften von Genexpressionsdaten, wie bei Thieffry et al. (1998). Verfahren, denen nur an einer mathematischen Beschreibung der gemessenen Daten ohne ein zugrundeliegendes Modell der Genexpression gelegen ist, werden hier ebenfalls nicht untersucht (z.B. Raychaudhuri et al., 2000; Alter et al., 2000).

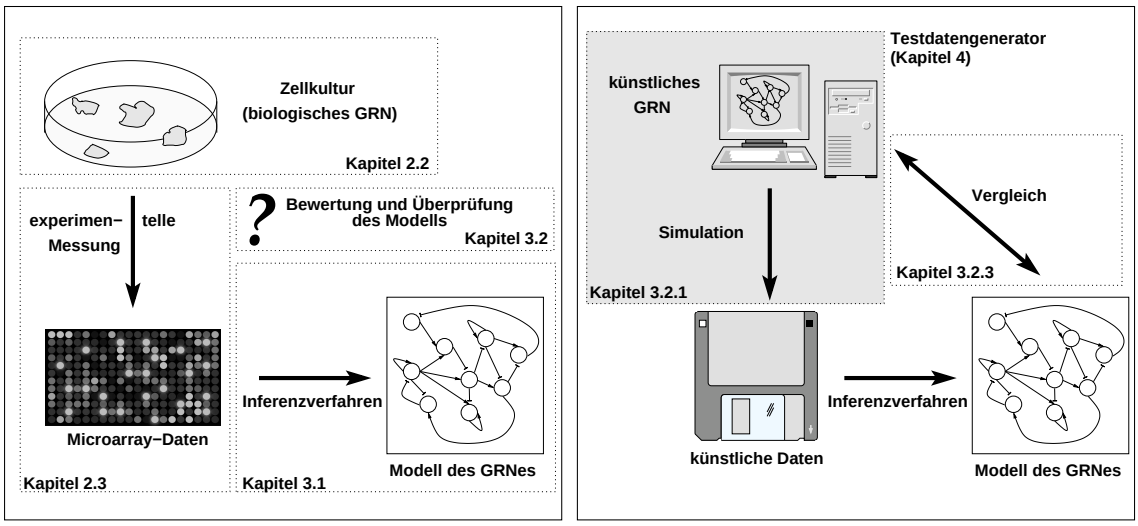

Abb. 3.3.: Inferenz eines biologischen Genregulationsnetzes (GRN) aus Microarray-Daten (links) und Bewertung eines Inferenzverfahrens durch ein künstliches Genregulationsnetz (rechts). Die gepunkteten Rahmen kennzeichnen die entsprechenden theoretischen Abschnitte dieser Arbeit, der grau unterlegte Rahmen verweist auf das Ziel des hier entwickelten Testdatengenerators.

Nach einer Klassifikation der Inferenzverfahren (Abschnitt 3.1 und Tabelle 3.1) wird in Abschnitt 3.2 beschrieben, wie die Ergebnisse von Inferenzverfahren bewertet werden können. Dabei liegt der Schwerpunkt auf Testdaten, die mit Hilfe künstlicher Genregulationsnetze gewonnen werden. In Abbildung 3.3 ist der Ablauf bei der Inferenz von biologischen Genregulationsnetzen sowie bei der Bewertung von Inferenzverfahren mit künstlichen Genregulationsnetzen dargestellt.

3.1. Klassifikation der Inferenzverfahren

Es existieren zahlreiche Methoden, durch die versucht wird, aus experimentellen Daten auf das zugrundeliegende Genregulationsnetz zu schließen. Diese Methoden unterscheiden sich sowohl in der Art der verwendeten Ausgangsdaten als auch in den zugrunde liegenden Modellen und Verfahren. In den folgenden Abschnitten wird eine Klassifikation der Inferenzmethoden entlang unterschiedlicher Merkmale vorgenommen. Dabei werden die grundlegenden Charakteristika der einzelnen Klassen besprochen und einige markante Literatur-Beispiele zu jeder Klasse gegeben. Eine Einordnung aller referierten Methoden erfolgt in einer Übersicht (Tabelle 3.1). Die **Klassen** (bzw. Unterklassen) sind im folgenden Text hervorgehoben, die in der Tabelle 3.1 benutzten Abkürzungen sind in Klammern angegeben. Auf eine detaillierte Besprechung der einzelnen Methoden wird verzichtet.

Eine Klassifikation der Modelle, die zur Inferenz von Genregulationsnetzen benutzt werden, nimmt de Jong (2002) vor. Dutilh und Hogeweg (1999) betrachten nur boolesche Modelle, Differenzialgleichungssysteme und Gewichtsmatrizen. Letztere repräsentieren jedoch auch Differenzialgleichungen, sodass die Einteilung eigentlich nur in zwei Klassen erfolgt. D'haeseleer et al. (2000) klassifiziert Inferenzverfahren nach einer Reihe von Merkmalen, unterteilt dabei jedoch weniger fein, als die hier vorgestellte Klassifikation.

3.1.1. Art der verwendeten experimentellen Daten

Von entscheidender Bedeutung für ein Inferenzverfahren ist die Art der Ausgangsinformation. Die wichtigste Quelle für die Inferenzverfahren für Genregulationsnetze ist zur Zeit die Bestimmung der Genaktivität über die experimentelle Messung der transkribierten mRNS-Menge. Fast alle Verfahren stützen sich ausschließlich auf diese Art von Information. Für die erfolgreiche Aufklärung der Regulationsvorgänge bei der Genexpression sollten jedoch sowohl vorhandenes biologisches Wissen (Datenbanken) als auch generelle Prinzipien biologischer Systeme (theoretische Biologie) in die Modellierung integriert werden (D'haeseleer, 2000, S. 132ff). Neben der Messung der mRNS-Konzentration ist es grundsätzlich auch möglich, die Konzentrationen von Proteinen und anderen Metaboliten zu messen. Einige Verfahren, wie das von Chen et al. (1999b), die explizit auch Proteinkonzentrationen modellieren, können diese zusätzliche Information ausnutzen. Eine Kombination von den hier vorgestellten Inferenzverfahren mit Methoden, die Genregulation auf der Sequenzebene untersuchen (z.B. Tavazoie et al., 1999), erscheint sinnvoll. Erste Schritte in dieser Richtung sind z.B. in der Arbeit von Davidson et al. (2002) zu sehen.

Bei der Messung der Genaktivität mittels mRNS-Konzentration gibt es grundsätzlich zwei verschiedene Vorgehensweisen, die großen Einfluss auf die Art der zu gewinnenden Information haben. Zum einen werden die zeitlichen Verläufe der mRNS-Konzentrationen der Gene verfolgt (**Zeitserien**, **ZS**). Mit Hilfe dieser Daten kann das dynamische Verhalten bei der Genexpression beobachtet werden. Demgegenüber stehen zum anderen Messungen der mRNS-Konzentrationen der Gene während eines relativ stabilen **Gleichgewichtszustands** (**GG**) (*Steady-State*) der Zelle.

Eine weitere Informationsquelle ist die Durchführung von **Störexperimenten** (**SE**). Dabei werden die mRNS-Konzentrationen der Gene unter verschiedenen Bedingungen gemessen. So ist es z.B. möglich, einzelne Gene genetisch dauerhaft zu deaktivieren (**single Deletion**, **sD**) oder zu verstärken (**Overexpression**, **sO**). Diese Art von Störung wird meist zusammen mit der *Steady-State*-Messung verwandt. Die mutierte Zellkultur wird

bis zum Gleichgewichtszustand entwickelt und anschließend die Messung vorgenommen. Als Referenz dient die Zellkultur im Normalzustand, die als Wildtyp bezeichnet wird. Damit kann man den Einfluss des veränderten Gens auf das restliche Genregulationsnetz beobachten. Auch **multiple** (**m**) genetische Störungen sind möglich, jedoch sinkt damit die Wahrscheinlichkeit, lebensfähige Zellen und damit Messwerte zu erhalten. Man kann bei diesen genetischen Störungen ebenfalls den Zeitverlauf der Genaktivitäten messen. Häufiger erfolgt die Messung von Zeitserien jedoch nach zeitlich begrenzter **Zugabe von Substanzen** (**Z**). So können mRNS oder Proteine zugegeben werden, um die Aktivität bestimmter Gene zu simulieren. Mittels Zugabe von RNS kann die komplementäre mRNS gebunden und eine Translation verhindert werden. Dies entspricht einem inaktiven Gen im Genregulationsnetz. Durch Zugabe von Signalstoffen der Zelle können bestimmte Zellereignisse vorgetäuscht werden.

3.1.2. Zugrundeliegender Modelltyp

Um aus den Genexpressionsdaten Aussagen über die Regulation der Genexpression treffen zu können, muss festgelegt sein, welcher Art diese Aussagen sein sollen. Es wird ein Modell der Vorgänge bei der Genexpression benötigt, welches an die experimentellen Daten angepasst werden soll. Einen vergleichenden Literaturüberblick über die Modellierung der Genregulation geben Smolen et al. (2000), einige allgemeine Betrachtungen zu Genregulationsmodellen gibt Reil (2000).

Grundsätzlich gibt es drei Ansätze, ein Genregulationsnetz zu modellieren: In **diskreten** Modellen werden die Änderungen der Genaktivitäten in diskreten Zeitschritten beschrieben. Die Genaktivitäten selbst werden ebenfalls als diskrete Werte ausgedrückt. Dadurch wird die rechentechnische Verarbeitung begünstigt. Demgegenüber stehen **kontinuierliche** Modelle, bei denen die Genaktivitäten mit Hilfe von Differenzialgleichungen ausgedrückt werden. Das gesamte Genregulationsnetz wird somit durch ein Differenzialgleichungssystem beschrieben, für die Lösung sind numerische oder Optimierungsverfahren notwendig. Weder die diskreten noch die kontinuierlichen Modelle berücksichtigen die stochastische Natur von chemischen Prozessen. Dies ist bei hohen Konzentrationen eine gute Näherung. Bei der Genregulation spielen aber auch Substanzen eine Rolle, von denen nur wenige Moleküle pro Zelle vorhanden sind (D'haeseleer, 2000, S. 43). Um dem gerecht zu werden, benutzt man **probabilistische** Modelle.

Diskrete Modelle

Die am weitesten verbreiteten diskreten Modelle sind sicherlich die **booleschen Netzwerke** (**BN**), deren Anwendung auf diese Domäne schon einige Jahrzehnte zurückreicht (Kauffman, 1969). Nach D'haeseleer et al. (1998) ist diese starke Vereinfachung, die Annahme von zwei Zuständen für die Aktivität eines Gens, für einige Gene eine gute Näherung: Zwischen der Grundaktivität und der von Regulatoren induzierten Aktivität liegen mehrere Größenordnungen (S. 6). Die Knoten im Netz repräsentieren die Gene und tragen einen binären Zustand: „ON“ (1) steht für ein aktives, „OFF“ (0) für ein inaktives Gen. Zwischen Knoten gibt es gerichtete Kanten, die den Einfluss zwischen den entsprechenden Genen symbolisieren. Jedem Gen mit k eingehenden Kanten ist eine k-stellige boolsche Funktion zugeordnet. Der Zustand des Gens wird in Abhängigkeit von den Zuständen der einwirkenden Gene mittels dieser Funktion berechnet. Bei **synchronen** booleschen Netzen erfolgt in diskreten Zeitschritten eine parallele Neuberechnung der Zustände aller Gene aus den Zuständen aller Gene im vorangegangenen Zeitpunkt. Solche synchronen booleschen

Netze weisen eine Reihe von interessanten Eigenschaften auf, mit denen sich z.B. Kauffman (1993) ausführlich beschäftigt hat. Oft werden die synchronen booleschen Netze auch einfach als boolesche Netze bezeichnet. Es ist jedoch auch möglich, die Neuberechnung der Zustände nicht gleichzeitig durchzuführen (**asynchron**). Diese Art boolescher Netze wird z.B. von Akutsu et al. (1998a) und Akutsu et al. (1998b) verwendet, die dynamischen Eigenschaften von asynchronen Netzen mit Rückkopplungen werden von Thieffry und Thomas (1998) untersucht. Von booleschen Netzen existieren eine Reihe von eingeschränkten und verallgemeinerten Varianten. So gibt es Beschränkungen im zulässigen Eingangsgrad (Kauffman, 1993; Akutsu et al., 1998b,a) ($\mathbf{k} \leq \mathbf{k_{max}}$), in der Art der verwendeten booleschen Funktionen (Kauffman, 1993) und Verbote von Rückkopplungen (Ideker et al., 2000).

Eine mögliche Erweiterung stellt die Benutzung von drei Zuständen dar: normal-, unter- und überexprimiert. Dies erlaubt eine bessere Modellierung von Delitions- und Überexpremierungsexperimenten. Solche **qualitativen Netzwerke** (**QN**) wurden z.B. von Akutsu et al. (2000) zur Modellierung eingesetzt: Die Gene werden wiederum durch Knoten in einem gerichteten Graphen repräsentiert, zwischen denen zwei Arten von Kanten (hemmend oder aktivierend) existieren können. Anstatt mit numerischen Werten für die Expressionswerte zu arbeiten, wird nur qualitativ beschrieben, welchen Einfluss die Expression eines Gens auf die Expression eines anderen Gens besitzt. Akutsu et al. (2000) weisen darauf hin, dass der Sinn nicht in Simulationen, sondern in der Repräsentation von biologischem Wissen liegt. Sie stellen ein Verfahren vor, mit dessen Hilfe ein qualitatives Netzwerk aus Genexpressionsdaten geschlossen werden kann. Die von Kyoda et al. (2000) bzw. Onami et al. (2000) zur Modellbeschreibung verwendete Interaktionsmatrix kann auch als qualitatives Netzwerk angesehen werden.

Eine besondere Form von qualitativen Netzen sind die von Maki et al. (2001) verwendeten ***multi-level Digraphs***: Bei den Knoten handelt es sich um Gene oder um Gruppen von Genen (*multilevel*). Zwischen diesen gibt es gerichteten Kanten, die einen Einfluss repräsentieren.

Kontinuierliche Modelle

Eine weitere Möglichkeit, die Vorgänge in Genregulationsnetzen zu modellieren, besteht in der Beschreibung der Reaktionskinetik der Genexpression. Dafür werden die zeitlichen Änderungsraten der Konzentrationen der betrachteten Substanzen mit Hilfe von Differenzialgleichungen angegeben. Das Gesamtsystem (das Genregulationsnetz) wird somit durch eine Menge von Differenzialgleichungen (ein **Differenzialgleichungssystem**, **DGL**) repräsentiert. Auf eine Darstellung der Reaktionskinetik der Genexpression soll hier verzichtet werden. Im Folgenden seien nur die wichtigsten Differenzialgleichungen angeführt, die zur Modellierung von Genregulationsnetzen verwendet werden. Eine detaillierte Einteilung und Beschreibung der verwendeten kontinuierlichen Modelle gibt de Jong (2002), eine Klassifikation nehmen Wessels et al. (2001) vor.

Die einfachste Form stellen **lineare Differenzialgleichungssysteme** dar, bei denen die Änderungsrate der Konzentration einer Substanz linear von den Konzentrationen regulierender Substanzen abhängt:

$$\frac{dy_i}{dt} = \sum_j w_{ji} y_j + b_i, \tag{3.1}$$

wobei y_i die Konzentration der Substanz i bezeichnet, w_{ji} repräsentiert den Einfluss (Gewicht) der Substanz j auf die Änderungsrate der Konzentration von Substanz i und b_i

ist die Änderungsrate von i in Abwesenheit jedes regulierenden Einflusses (D'HAESELEER, 2000, S. 6). Da die Einflüsse der regulierenden Substanzen addiert werden, bezeichnet D'HAESELEER ET AL. (2000) sie als „additive regulation models“ (S. 9).

Diese starke Vereinfachung ist biochemisch recht unplausibel, da Konzentrationen in diesem Modell beliebig groß oder auch negativ werden können (D'HAESELEER, 2000, S. 7). Um dies zu verhindern, wird oft statt einer linearen Abhängigkeit eine Funktion verwendet, die nach unten und oben beschränkt ist, entweder durch konstante Werte (**„logoid“**, DE JONG, 2002, S. 78) oder asymptotisch (**sigmoid**). Dies ist beispielhaft in Abbildung 3.4 dargestellt. Es gibt mehrere Möglichkeiten von sigmoiden Funktionen. Häufig wird die Funktion $f(x) = \frac{a}{1+e^{-x}}$ verwendet, wobei a einen Verstärkungsfaktor angibt, der verschieden gewählt wird (ANDO UND IBA, 2001; WAHDE UND HERTZ, 2000; WEAVER ET AL., 1999). D'HAESELEER (2000) schlägt auch die Funktion $f(x) = \tanh(x)$ vor. Eine weitere Variante wird von KYODA ET AL. (2000) zur Simulation der Genexpression benutzt: $f(x) = \frac{1}{2}\left(x/\sqrt{x^2+1}+1\right)$. Logoide Funktionen wurden z.B. von MOROHASHI UND KITANO (1999), sigmoide Funktionen, außer den oben erwähnten, z.B. noch von MJOLSNESS ET AL. (1999) verwendet.

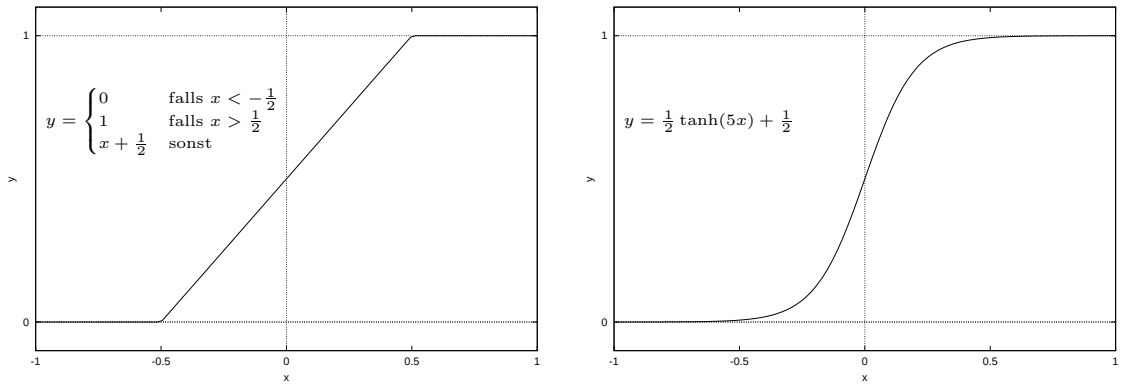

Abb. 3.4.: Beispiele für beschränkte Funktionen: logoide Funktion (links) und sigmoide Funktion (rechts).

Eine Möglichkeit, zwei fundamentale Eigenschaften biochemischer Systeme sicherzustellen, bieten **S-Systeme** (**S**): Das „S“ steht für Sättigung (*Saturation*) und Synergismus, welche den S-Systemen inhärent sind (VOIT, 2000, S. 51). Die grundsätzliche Form dieser Differenzialgleichungen ist die Differenz zwischen Produktion (erster Term) und Abbau (zweiter Term):

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{n+m} x_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} x_j^{h_{ij}} \qquad \text{für } i = 1, \ldots, n, \tag{3.2}$$

wobei $x_1, \ldots, x_n$ die vom System abhängigen und $x_{n+1}, \ldots, x_{n+m}$ die unabhängigen Konzentrationsvariablen sind. $\alpha_i \geq 0$ ist die Ratenkonstante der Produktion und $\beta_i \geq 0$ ist die Ratenkonstante des Abbaus von x_i. Die Exponenten $g_{ij}, h_{ij} \in \mathbb{R}$ sind die kinetischen Ordnungen der Produktion bzw. des Abbaus von Substanz i in Bezug auf Substanz j. (VOIT, 2000). Durch die Produktform kann es zu einer synergetischen Verstärkung der Einflüsse von Substanzen auf die Konzentrationsänderung von i kommen. Bei geeigneter Wahl der Parameter können sich die Produktions- und Abbaurate einer Substanz angleichen, wodurch die Konzentration dieser Substanz einer Sättigungskonzentration entgegenstrebt (*Saturation*).

Zur Modellierung von Genregulationsnetzen wurden die S-Systeme z.B. von Akutsu et al. (2000) und Maki et al. (2001) benutzt. Für die Darstellung als Produkte von Potenzen gibt es nach Savageau (1998) sowohl theoretische als auch empirische Rechtfertigungsgründe (S. 7).

Kontinuierliche Modelle werden häufig um zwei Merkmale erweitert: Zum einen werden **externe Einflüsse** (**ext**) berücksichtigt, indem zusätzliche Variablen eingeführt werden, die analog den Konzentrationsvariablen auf der rechten Seite von Differenzialgleichungen verwendet werden können. Die andere Erweiterung ist die Berücksichtigung des **Zerfalls** (*Decay*) von biochemischen Substanzen (**D**). Da Zerfallsreaktionen kinetisch durch Reaktionen erster Ordnung beschrieben werden können, erfolgt die Berücksichtigung durch Subtraktion eines Ausdrucks der Form $D_i y_i$ (D'haeseleer, 2000), wobei D_i die Zerfallsrate der Substanz i ist. Der Gesamtausdruck für die Transkription der Substanz i lautet damit bei D'haeseleer (2000):

$$\frac{dy_i}{dt} = A_i S\Big(\sum_j w_{ji} y_j + b_i\Big) - D_i y_i, \tag{3.3}$$

wobei S eine Sigmoidfunktion und A_i die maximale Transkriptionsrate von i darstellt. Sowohl der Gewebetyp, als auch die externe Zugabe einer Substanz werden dabei durch zusätzliche y-Variablen repräsentiert, deren Einfluss auf die Transkription einzelner Gene durch entsprechende Gewichte beschrieben wird.

Probabilistische Modelle

Will man der stochastischen Natur der Vorgänge bei der Genexpression gerecht werden, sind die bisher besprochenen deterministischen Modelle ungeeignet. Ereignisse, wie z.B. der Beginn der Transkription eines bestimmten Gens, müssen vielmehr als zufällig angesehen werden. Das System kann durch eine Menge von Zufallsvariablen und Wahrscheinlichkeitsverteilungen für diese Variablen beschrieben werden. Die Zufallsvariablen können z.B. für die Konzentrationen von Substanzen oder für bestimmte Ereignisse stehen. Einflüsse zwischen Genen erscheinen in Form von bedingten (Un)Abhängigkeiten zwischen entsprechenden Zufallsvariablen.

Eine verbreitete Form, die bedingten Unabhängigkeiten zwischen Zufallsvariablen darzustellen, sind die **Bayes-Netze** (**BayN**), in die z.B. Charniak (1991) eine Einführung gibt. Ein Bayes-Netz besteht aus:

1. einem gerichteten, azyklischen Graphen, dessen Knoten Zufallsvariablen repräsentieren und dessen Kanten direkte Einflüsse zwischen diesen Zufallsvariablen darstellen und

2. aus der Wahrscheinlichkeitsverteilung für alle Zufallsvariablen: Für Knoten, die keine eingehende Kante besitzen, müssen a priori Wahrscheinlichkeiten angegeben werden. Besitzt ein Knoten eingehende Kanten, so ist ihm eine bedingte Wahrscheinlichkeit (bedingt durch die Werte der Zufallsvariablen, die einen direkten Einfluss auf ihn haben) zugeordnet.

Die Zufallsvariablen in einem Netz können entweder diskrete oder kontinuierliche Werte annehmen. Davon abhängig erfolgt die Darstellung der bedingten Wahrscheinlichkeiten: Im Fall diskreter Variablen kann für jede mögliche Kombination von Werten der Variablen, die einen direkten Einfluss ausüben, der Wert der bedingten Wahrscheinlichkeit angegeben

werden (Wertetabelle). Eine solche diskrete Verteilung wird in Verallgemeinerung der Binominalverteilung (bei der die Zufallsvariable nur zwei mögliche Werte besitzen können) als **multinominale Verteilung** (**multiV**) bezeichnet. Im kontinuierlichen Fall ist die Angabe einer Wertetabelle nicht möglich. Friedman et al. (2000) verwenden dafür eine **lineare Gaussverteilung** (**GV**), d.h. die bedingte Wahrscheinlichkeit ist normalverteilt um einen Wert, der linear von den beeinflussenden Variablen abhängt. Auch Imoto et al. (2002) verwenden eine Normalverteilung, jedoch hängt der Mittelwert **nichtlinear** (**nichtlin. GV**) von den Werten der beeinflussenden Variablen ab (sie bezeichnen dies als „nonparametric regression model with Gaussian noise“, Imoto et al., 2002, S. 3).

Die Modellierung von Genregulationsnetzen mit Bayes-Netzen besitzt neben der Möglichkeit, stochastische Vorgänge zu berücksichtigen, weitere Vorteile. So können verrauschte Daten und unvollständiges Wissen gut behandelt werden (de Jong, 2002). Störexperimente können durch Setzen von a priori Wahrscheinlichkeiten für die entsprechenden Variablen modelliert werden (Pe'er et al., 2001; Yoo et al., 2002). Einen Nachteil stellt die statische Natur der Netze dar: Die dynamischen Vorgänge in einem Genregulationsnetz können nicht explizit repräsentiert werden (de Jong, 2002). Obwohl Erweiterungen der Bayes-Netze zur Modellierung dynamischer Prozesse existieren, werden sie zur Zeit noch nicht in dieser Domäne eingesetzt.

Eine mögliche Erweiterung stellt die Einführung **beschrifteter Kanten** (**beschr.**) dar, wie sie z.B. von Hartemink et al. (2001) zur Modellierung von Genregulationsnetzen verwendet wird: Eine Kante kann einen beliebigen, einen positiven, einen negativen oder einen unbekannten Einfluss besitzen. Bei Pe'er et al. (2001) wird zwischen aktivierenden und inhibierenden Einflüssen unterschieden.

Toh und Horimoto (2002) verwenden einen den Bayes-Netzen ähnlichen Ansatz zur Modellierung von bedingten Abhängigkeiten zwischen Paaren von Genen. Im Gegensatz zu Bayes-Netzen handelt es sich bei ihren **Unabhängigkeitsgraphen** („independence graph“, Toh und Horimoto, 2002, S. 289) jedoch um ungerichtete Graphen: eine Kante zwischen zwei Knoten (Genen) symbolisiert einen Einfluss *zwischen* den beiden Genen.

3.1.3. Art des verwendeten Verfahrens

Um aus den experimentellen Daten ein Genregulationsnetz entsprechend dem gewählten Modell abzuleiten, benötigt man ein geeignetes Inferenzverfahren. Dabei stellt die Fragestellung „Mit welcher konkreten Ausprägung des Modells können die gemessenen Daten am besten erklärt werden?“ eine typische Aufgabe des *Reverse Engineering* dar. Was hierbei unter einer Erklärung zu verstehen ist, hängt vom zugrundeliegenden Modell ab und wird an entsprechender Stelle erläutert (siehe unten). Für die Lösung existieren eine Fülle von verschiedenartigen Verfahren, die zur Inferenz von Genregulationsnetzen eingesetzt werden können. Meist werden die Verfahren jedoch abgewandelt bzw. mehrere Verfahren kombiniert, was eine strikte Klassifikation erschwert. Aus diesem Grund erfolgt hier nur eine grobe Einteilung der Verfahren, die einzelnen Ausprägungen werden in der folgenden Übersicht nur genannt. Für Details sei auf die jeweilige Quelle verwiesen.

Die Wahl des Verfahrens hängt stark von der Art der zur Verfügung stehenden Daten und vom zugrundegelegten Modell ab (siehe 3.1.1 und 3.1.2). Darüber hinaus spielen auch die Qualität und Quantität der experimentellen Daten, das gewünschte Maß an Details des inferierten Netzes und die zur Verfügung stehenden Rechenressourcen eine entscheidende Rolle. Für die Begründungen der Methodenauswahl sei ebenfalls auf die jeweilige Quelle verwiesen.

Einige der besprochenen Verfahren suchen nach Kanten nicht zwischen Genen, sondern zwischen Genclustern. Dafür wird in einem vorangehenden Schritt eine **Clusterung** (**C**) durchgeführt, bei der meist Gene mit ähnlichen Expressionsprofilen zusammengefasst werden. Anschließend erfolgt die Inferenz eines Netzes zwischen diesen Clustern (Mjolsness et al., 1999; Toh und Horimoto, 2002). Chen et al. (1999a) benutzen Clusterung, um die Anzahl der betrachteten Gene zu reduzieren: Gene, deren Expressionsprofile nahezu übereinstimmen, werden als identisch betrachtet. Es gibt darüber hinaus eine Vielzahl von Verfahren, deren Hauptziel die Clusterung von Genen ist. Diese liegen jedoch außerhalb dieser Untersuchung.

Verfahren für kontinuierliche Modelle

Bei kontinuierlichen Modellen werden die Vorgänge bei der Genexpression durch Differenzialgleichungen beschrieben. Einem Genregulationsnetz entspricht somit ein Differenzialgleichungssystem: Für jeden Knoten (Gen) beschreibt eine Differenzialgleichung die Änderungsrate der Konzentration des Primärtranskripts dieses Gens in Abhängigkeit der Konzentrationen der regulierenden Substanzen. Die regulierenden Substanzen werden durch eingehende Kanten von den Genen, die für ihr Entstehen verantwortlich sind, repräsentiert. Die Stärke eines regulierenden Einflusses kann durch ein Kantengewicht beschrieben werden. Neben diesen Kantengewichten enthalten die Differenzialgleichungen meist noch andere (kinetische) Parameter. Weitere Vorgänge, wie Translation oder Dimerisation, können durch zusätzliche Differenzialgleichungen beschrieben werden.

Wird ein kontinuierliches Modell verwendet, besteht die Aufgabe der Inferenz darin, einen Satz von Parametern für das Differenzialgleichungssystem zu suchen, mit dessen Hilfe sich die experimentellen Daten (meist Zeitserien) approximieren lassen. Meist wird dabei von einem vollständig verbundenen Netz ausgegangen, d.h. jedes Gen kann potenziell jedes andere Gen regulieren. Es wird versucht, sowohl die Kantengewichte als auch die kinetischen Parameter zu ermitteln. Regulierende Einflüsse im realen Genregulationsnetz sollten bei diesem Verfahren möglichst deutlich von Null verschiedene Kantengewichte erhalten, während ein real nicht vorhandener Einfluss zu einer Kante mit annähernd verschwindendem Gewicht führen sollte. Bei einfachen Modellen, wie linearen Differenzialgleichungen, ist eventuell eine **algebraische Lösung** (**ALG**) möglich (z.B. D'haeseleer, 2000). Die gleichzeitige Bestimmung von Struktur und Dynamik des Netzes führt jedoch zu einer riesigen Anzahl anzupassender Parameter, wodurch das System schnell unterbestimmt wird, wenn nicht genug Messwerte zur Verfügung stehen. In diesem Fall sind Näherungslösungen notwendig, entweder durch **numerische Verfahren** (**NUM**) oder **Optimierungsmethoden** (**OPT**). Zu den numerischen Verfahren zählen Differenzialgleichungslöser (z.B. Chen et al., 1999b) und *Singular Value Decomposition* (z.B. Weaver et al., 1999; Yeung et al., 2002; Holter et al., 2001). Optimierungsverfahren sind z.B. Gradientenabstiegsverfahren (D'haeseleer, 2000), evolutionäre (insbesondere genetische) Algorithmen (Ando und Iba, 2001; Maki et al., 2001; Morohashi und Kitano, 1999; Wahde und Hertz, 2000) und *Simulated Annealing* (Mjolsness et al., 1999).

Bei Morohashi und Kitano (1999) handelt es sich um ein mehrstufiges Verfahren: Im ersten Schritt werden Kandidaten für Genregulationsnetze geraten (**Sampling**). Anschließend werden die Parameter dieser Kandidatennetze mittels eines genetischen Algorithmus an die Zeitserien angepasst. Danach erfolgt mit Hilfe von Störexperimenten eine weitere Selektion der Netze (**Screening**).

Verfahren für diskrete Modelle

Bei der Darstellung eines Genregulationsnetzes als boolesches Netz (V, E) wird jedes Gen durch einen Knoten $v \in V$ repräsentiert, der zwei Zustände annehmen kann. Einem Knoten v mit k eingehenden Kanten $(u_1, v), \ldots, (u_k, v) \in E$ ist eine k-stellige boolesche Funktion $f_v : \mathbb{B}^k \to \mathbb{B}$ zugeordnet, die in Abhängigkeit der Zustände von $u_1, \ldots, u_k$ den Zustand des Knoten v berechnet. Bei synchronen booleschen Netzen erfolgt die Neuberechnung der Zustände aller Knoten in diskreten Zeitschritten, der Zustand zum Zeitpunkt $t + 1$ wird aus den Zuständen aller Knoten zum Zeitpunkt t mit Hilfe der booleschen Funktionen berechnet. Die Inferenz von synchronen booleschen Netzen kann aus Zeitserien erfolgen. Aus zwei aufeinanderfolgenden Messpunkten der Zeitserie kann ein Paar Zustandsvektoren der Gene gewonnen werden. Ein solches Paar wird als Ein-/Ausgabepaar bezeichnet (Akutsu et al., 1999, 2000). Ein synchrones boolesches Netz erklärt eine Menge solcher Ein-/Ausgabepaare, falls es in einem Zeitschritt vom Eingabezustand zum Ausgabezustand gelangt. Genügend viele Ein-/Ausgabepaare des betrachteten Netzes und beschränkte Stelligkeit der booleschen Funktionen vorausgesetzt, ist eine erschöpfende **Suche** im Raum der möglichen booleschen Funktionen zur Netzinferenz möglich (Akutsu et al., 2000; Akutsu et al., 1999). Eine mehr systematische Suche im Raum der booleschen Funktionen führen Liang et al. (1998) durch: Die Suche erfolgt in Richtung steigenden Eingangsgrads der Funktionen, die Bewertung der Funktionen erfolgt mit Mitteln der Informationstheorie (*mutual information analysis*).

Akutsu et al. (1998a,b) verwenden für die Inferenz von asynchronen booleschen Netzen Störexperimente (multiple Löschungen und Überexprimierungen von Genen). Jedes dieser Experimente liefert einen Zustandsvektor. Die Knoten der gestörten Gene werden im booleschen Netz auf dem entsprechenden Zustand festgehalten (0 bei Löschung und 1 bei Überexprimierung). Besitzt das so modifizierte asynchrone boolesche Netz einen Folgezustand, der dem Zustandvektor der experimentellen Messung entspricht, so wird das Netz als konsistent mit einem Störexperiment angesehen. Der Folgezustand ist bei asynchronen booleschen Netzen von der Reihenfolge der Neuberechnung der Zustände der einzelnen Knoten abhängig und somit im Allgemeinen nicht eindeutig bestimmt. Akutsu et al. (1998b) fordern für Konsistenz nur, dass es **einen** solchen Folgezustand gibt. Ein asynchrones boolesches Netz ist eine Erklärung einer Menge von Störexperimenten, falls es mit jedem dieser Experimente in dem genannten Sinne konsistent ist. Die Inferenz erfolgt bei Akutsu et al. (1998b,a) ebenfalls durch Suche nach geeigneten booleschen Funktionen.

Für die Inferenz von azyklischen booleschen Netzen verwenden Ideker et al. (2000) ebenfalls Störexperimente. Hier werden jedoch nicht, wie bei der Gruppe um Akutsu et al. Folgezustände betrachtet, sondern ausschließlich lokale Konsistenz: Werden alle Knoten mit den Zuständen eines Experiments belegt, müssen die booleschen Funktionen der Knoten wahre Aussagen darstellen. Ein azyklisches boolesches Netz erklärt eine Menge von Störexperimenten, falls es für jedes Experiment in diesem Sinne korrekt ist. Um ein solches Netz aus einer Menge von Störexperimenten zu schließen, führen Ideker et al. (2000) für jedes Gen einen paarweisen **Vergleich** (**VGL**) der Zustände des Gens in je zwei Experimenten durch. Ändert sich der Zustand, wird die Menge derjenigen Gene bestimmt, deren Zustand sich ebenfalls zwischen diesen Experimenten ändert. Es wird davon ausgegangen, dass mindestens eines dieser Gene für die Zustandsänderung des betrachteten Gens verantwortlich ist. Dieses Vorgehen wird für alle Experimentpaare wiederholt und so eine kleinste Mengen von Genen bestimmt, die alle beobachteten Änderungen des betrachteten Gens verantwortet haben könnten. Ein Suchverfahren bestimmt für jede dieser Mengen eine boolesche Funktion, die von den Genen in der Menge abhängig und lokal konsistent mit den Experimenten ist.

In den qualitativen Netzwerken von Kyoda et al. (2000) werden die Gene ebenfalls durch

Knoten repräsentiert. Eine Kante stellt einen aktivierenden oder inhibierenden Einfluss von einem Gen auf ein anderes dar. Zur Inferenz werden einfache Löschungs- oder Überexprimierungsexperimente verwendet. Das Verfahren beruht auf paarweisem Vergleich der Aktivität eines Gens im Wildtyp und in einem Störexperiment. Ändert sich diese Aktivität, so wird dies durch eine Kante vom gelöschten bzw. überexprimierten Gen zum betrachteten Gen dargestellt. Diese Kante ist aktivierend, falls die Aktivitätsänderung der Störung entspricht (Anstieg bei Überexprimierung bzw. Abfall bei Löschung), anderenfalls inhibierend. Unter der Annahme minimaler Kantenzahl wird ein qualitatives Netzwerk gesucht, in dem für jede Kante ein Experiment existiert, das diese Kante in der eben angeführten Weise rechtfertigt. Für die Minimierung der Kantenanzahl wird ein **Graphenalgorithmus** (**GR**) verwendet.

Maki et al. (2001) stellen in ihren *multi-level Digraphen* durch gerichtete Kanten einen Einfluss zwischen zwei Genen bzw. Gengruppen (siehe unten) dar. Die Kanten werden durch Vergleich der Aktivität eines Gens im Wildtyp und in einem einfachen Störexperiment gefunden, in ähnlicher Art wie bei Kyoda et al. (2000). Ein Problem kann durch die gleichzeitige Expression (co-Expression) von Genen entstehen: Die Beeinflussung zwischen diesen Genen ist dann eventuell wechselseitig, eine eindeutige, gerichtete Kante nicht mehr möglich. Maki et al. (2001) fassen solche Gene zu Äquivalenzklassen zusammen und behandeln diese Klasse im weiteren Ablauf anstelle der einzelnen Gene.

Eine Repräsentation der zeitlichen Abhängigkeiten zwischen den Aktivitäten von Genen durch qualitative Netzwerke (Chen et al., 1999a; Akutsu et al., 2000) bzw. durch Unabhängigkeitsgraphen (Toh und Horimoto, 2002) kann aus Zeitserien gewonnen werden: Durch Vergleich der Expressionsprofile zweier Gene (**zeitVGL**) können zeitliche Abhängigkeiten gefunden werden (Chen et al., 1999a). Dies kann auch mit statistischen Mitteln der **Korrelationsanalyse** (**KORR**) erfolgen. Durch ein solches Verfahren konstruieren Toh und Horimoto (2002) ihren Unabhängigkeitsgraphen. Akutsu et al. (2000) verwenden eine weitere statistische Methode zur Inferenz von qualitativen Netzwerken: ***Linear Programming*** (**LP**) ist eine Art lineare Regression für Ungleichungssysteme.

Verfahren für Bayes-Netze

Die Inferenz eines Bayes-Netzes aus experimentellen Daten kann mit Hilfe von **Lernverfahren** für Bayes-Netze (**bayLV**) erfolgen. Dafür wird jedes Netz hinsichtlich seiner Eignung zur Repräsentation der experimentellen Beobachtungen mittels einer Bewertungsfunktion (*Scoring Function*) beurteilt. So wird aus der Lernaufgabe ein Optimierungsproblem, das z.B. mit Hilfe heuristischer Suche im Raum der möglichen Netze gelöst werden kann (Friedman et al., 2000). Die Lernverfahren für Bayes-Netze, die zur Inferenz von Genregulationsnetzen eingesetzt werden, unterscheiden sich sowohl in der Bewertungsfunktion, als auch im verwendeten Optimierungsverfahren. Für Details sei auf die jeweilige Quelle verwiesen.

Referenz	Daten	Modelltyp	Verfahren	Bemerkungen
Akutsu et al. (1998b,a)	SE: mD, mO (GG)	BN (asynchron) [$k \leq 2$, AND oder OR]	Suche ↕ EXP	
Akutsu et al. (1999)	ZS	BN [$k \leq k_{max}$]	Suche	
Akutsu et al. (2000)	ZS	1. BN [$k \leq k_{max}$] 2. QN 3. DGL: S	1. Suche 2. LP 3. modifiziertes LP	verrauschte Daten berücksichtigt, (3.) erzeugt qualitatives DGL
Ando und Iba (2001)	1. ZS 2. SE	DGL: sig(Σ)	1. OPT (GA) 2. VGL ↓ OPT (wie 1.)	GA arbeitet global bzw. erst lokal und anschließend global
Chen et al. (1999b)	ZS	DGL: $\Sigma - D$	NUM (FTSS, MWSLE)	Transkription, Translation und Verzögerungen berücksichtigt
Chen et al. (1999a)	ZS	QN	zeitVGL ↓ OPT (SA)	Anzahl von Genen mit mindestens einem Eingang maximiert, Gen ist entweder Aktivator oder Inhibitor
D'haeseleer (2000)	SE: Z (ZS)	1. DGL: Σ 2. DGL: sig(Σ + ext) $- D$	1. ALG 2. OPT (NN-LV)	siehe auch D'haeseleer und Fuhrman (1999), D'haeseleer et al. (1999)
Friedman et al. (2000)	ZS	BayN (multiV,GV)	bayLV	Vertrauenswerte für gefundene Merkmale
Hartemink et al. (2001)	ZS	BayN (beschr.)	bayLV	
Holter et al. (2001)	ZS	DGL: Σ	NUM (SVD) ↓ OPT (SA)	charakteristische Modi
Ideker et al. (2000)	SE:* (GG)	BN (azyklisch)	VGL ↓ OPT (BB) ↓ EXP	minimale Kantenanzahl
Imoto et al. (2002)	ZS	BayN (nichtlin. GV)	bayLV	
Kyoda et al. (2000)	SE: sD (GG)	QN (trinäre Interaktionsmatrix)	VGL ↓ GR	siehe auch Onami et al. (2000), keine indirekten Kanten
Liang et al. (1998)	ZS	BN	Suche (MIA)	
Maki et al. (2001)	SE: sD,sO (ZS, GG)	*multi-level Digraph* ↓ DGL: S	VGL ↓ OPT (GA)	VGL erzeugt Äquivalenzklassen von Genen
Mjolsness et al. (1999)	ZS	DGL: sig(Σ) $- D$	C (EM) ↓ OPT (GA)	
Morohashi und Kitano (1999)	SE: D,O (ZS)	DGL: logo(Σ) $- D$	Sampling ↓ OPT (GA) ↓ Screening	siehe Seite 26

Tab. 3.1.: Klassifikation von Inferenzmethoden

Referenz	Daten	Modelltyp	Verfahren	Bemerkungen
Pe'er et al. (2001)	SE: * (ZS)	bayN (beschr.)	bayLV	detailliertere Inferenz „sichere“ Unternetze, Modellierung von Störungen
Toh und Horimoto (2002)	SE: * (ZS)	Unabhängigkeitsgraph	C (H) ↓ KORR (GGM)	
Wahde und Hertz (2000)	SE: * (ZS)	DGL: sig(Σ)	OPT (GA)	
Weaver et al. (1999)	ZS	DGL: sig(Σ + ext)	OPT (SVD, PIMP)	
Yeung et al. (2002)	SE: Z (ZS)	DGL: $\Sigma - D$	NUM (SVD) ↓ NUM (RR)	dünn besetzte Verbindungsmatrix
Yoo et al. (2002)	SE: sD (GG)	bayN	bayLV	versteckte Variablen, Modellierung von Störungen

Tab. 3.1.: Klassifikation von Inferenzmethoden, Fortsetzung

Hinweise zur Tabelle:
Die Tabelle enthält die Klassifikation von Methoden zur Inferenz von Genregulationsnetzen entsprechend der vorangegangenen Kapitel. Die erste Spalte enthält die Referenz. Die letzte Spalte enthält einige Bemerkungen, vor allem einschränkende Annahmen sowie zusätzliche Fähigkeiten der Verfahren und Hinweise auf weitere Referenzen. Die mittleren Spalten entsprechen den behandelten Klassifikationsmerkmalen. Eine Nummerierung innerhalb eines Eintrags weist auf alternative Ansätze hin, ein Pfeil auf die Hintereinanderausführung von Verfahren bzw. die Kombination von Modellen. Es folgt eine Auflistung der verwendeten Abkürzungen:

2. Spalte: Art der verwendeten **Daten**
- **ZS** Messung von Zeitserien
- **GG** Messung im Gleichgewicht (*Steady-State*)
- **SE** Störexperiment: **D**=Löschung (*Deletion*), **O**=Überexprimierung (*Overexpression*), **Z**=Zugabe von Substanzen, *****=verschiedenartige, **s**=einfache (*single*), **m**=multiple

3. Spalte: Zugrundeliegender **Modelltyp**
- **BN** boolesches Netzwerk: synchron (ohne Angabe) bzw. **asynchron**, Einschränkungen werden in eckigen Klammern [ ] angegeben - $\mathbf{k} \leq \mathbf{k_{max}}$=Eingangsgrad ist beschränkt (auch Zahlenangabe möglich), **AND** bzw. **OR**=boolesche Funktionen sind auf $\wedge$ bzw. $\vee$ beschränkt, die Eingänge können allerdings negiert sein
- **QN** qualitatives Netzwerk
- **DGL** Differentialgleichungssystem: **S**=S-System, **Σ**=gewichtete Summe der Eingänge, **sig**=sigmoide Funktion, **logo**=logoide Funktion, **−D**=Zerfall wird berücksichtigt
- **BayN** Bayes-Netz: **GV**=lineare Gaussverteilung, **nichtlin. GV**=nichtlineare Gaussverteilung, **multiV**=multinominale Verteilung, **beschr.**=beschriftete Kanten

4. Spalte: Art des verwendeten **Verfahren**s
- **OPT** Optimierungsverfahren: **GA**=genetischer Algorithmus, **SA**=*Simulated Annealing*, **NN-LV**=Lernverfahren für Neuronale Netze, **BB**=*Branch and Bound*
- **NUM** numerisches Verfahren: **MWSLE**=*Minimum Weight Solutions to Linear Equations*, **FTSS**=*Fourier Transform for Stable Systems*, **SVD**=*Singular Value Decomposition*, **PIMP**=Pseudoinverse nach Moore-Penrose, **RR**=robuste Regression
- **VGL** Verfahren, das auf Vergleich der Aktivitäten eines Gens unter verschiedenen Bedingungen basiert
- **zeitVGL** Vergleich der Zeitserien von Genen
- **MIA** *Mutal Information Analysis*
- **KORR** Korrelationsanalyse: **GGM**=*Graphical Gaussian Modelling*
- **EXP** Vorschlagen weiterer Experimente
- **GR** Graphenalgorithmus
- **LP** *Linear Programming*
- **C** vor der Inferenz wird eine Clusterung vorgenommen: **H**=hierarchische Clusterung, **EM**=*Expectation Maximization*

3.2. Bewertung von Inferenzmethoden

Im letzten Abschnitt wurden Methoden vorgestellt, mit deren Hilfe von Microarray-Daten auf das zugrundeliegende Genregulationsnetz geschlossen werden kann. Wie können die Ergebnisse solcher Methoden überprüft, wie kann deren Leistungsfähigkeit bewertet werden? Nach einer Übersicht über mögliche Ansätze zur Leistungsbewertung wird im Abschnitt 3.2.1 auf die Erzeugung künstlicher Genregulationsnetze näher eingegangen. Es schließt sich eine Kritik an den verwendeten Modellsystemen an (3.2.2) und der letzte Abschnitt (3.2.3) behandelt Maße, mit denen die Bewertung erfolgen kann.

Zum einen kann auf vorhandenes **biologisches Wissen** zurückgegriffen werden. Für einige Gene ausgewählter Organismen ist ein Teil des Regulationsmechanismus bekannt. Diesen Abgleich der Ergebnisse mit existierendem Wissen verwenden z.B. Yoo et al. (2002), Ideker et al. (2000), Toh und Horimoto (2002), D'haeseleer et al. (1999) und Friedman et al. (2000). Die rasche Entwicklung und zunehmende Verbreitung der Microarray-Technik in den letzten Jahren führt zu einer wahren Datenflut über die Genexpression vieler tausender, bislang nicht untersuchter Gene von immer mehr Organismen. Das bekannte biologische Wissen über die Regulation der Genexpression, das auf klassischem Wege gewonnen wurde, kann somit nur für einen kleinen Teil der Ergebnisse als Prüfstein dienen und stellt nur erste Hinweise auf die Leistungsfähigkeiten der Inferenzverfahren zur Verfügung.

Die natürlich vorkommenden Genregulationsnetze sind zum Großteil unbekannt. Zur Bewertung von Inferenzverfahren kann aber ein Modellsystem (ein **künstliches Genregulationsnetz**) konstruiert werden. Mit diesem werden die Vorgänge bei der Genexpression in der Natur nachgeahmt und so künstliche Genexpressionsdaten erzeugt. Das Ergebnis der Anwendung einer Inferenzmethode auf diese künstlichen Daten kann dann direkt mit dem bekannten Netz verglichen werden (Liang et al., 1998, S. 20).

Weiterhin kann überprüft werden, ob die aus Microarray-Daten inferierten Netze **biologisch plausibel** sind. Dies stellt zwar ebenfalls einen Rückgriff auf biologisches Wissen dar, jedoch nicht auf Detailwissen über die Regulation einzelner Gene, sondern auf ein allgemeines Verständnis der Vorgänge bei der Genexpression. So führen D'haeseleer et al. (1999) Eigenschaften von Genregulationsnetzen an, auf die sie ihre inferierten Netze überprüfen (siehe auch 2.2).

Wird ein kontinuierliches Modell verwendet, kann dessen **Vorhersagekraft** bewertet werden: Kann das Modell die gemessenen Daten reproduzieren? Die Ähnlichkeit zwischen den vom Modell erzeugten Daten und den Messdaten dient den meisten Optimierungsverfahren als Orientierungs- und Endkriterium. Zur Bewertung explizit benutzt wird dieses Kriterium z.B. von Holter et al. (2001) und Wessels et al. (2001).

Ein weitere Möglichkeit besteht darin, gefundenen Parametern **Vertrauenswerte** zuzuweisen: Der Einfluss von kleinen Störungen der Messwerte auf die Ergebnisse wird untersucht. Werten, die sich gegenüber solchen Störungen wenig empfindlich erweisen, wird mehr vertraut, als Werten, die eine eher chaotische Reaktion zeigen. Eine solche statistische Bewertung findet sich vor allem bei probabilistischen Netzwerkmodellen (*Bootstrap*-Methode: Toh und Horimoto, 2002 und Friedman et al., 2000, Messung der Signifikanz: Pe'er et al., 2001). D'haeseleer (2000) benutzt eine solche Analyse ebenfalls, um „robuste Parameter" zu finden.

In diesem Zusammenhang werden oft **Kontrollmodelle** verwendet: Bei D'haeseleer (2000) dienen völlig zufällig erzeugte Genexpressionsdaten als Kontrollmodell und Fried-

MAN ET AL. (2000) benutzen willkürlich permutierte Messdaten. Da in den so erhaltenen Daten keine echten Einflüsse vorhanden sind, ist zu erwarten, dass das Inferenzverfahren keine signifikanten Merkmale (Parameter mit hohem Vertrauenswert, siehe oben) findet.

Die von HASTY ET AL. (2001) beschriebene Methode zur **gentechnischen** Erzeugung von **synthetischen Genregulationsnetzen** stellt möglicherweise auch eine Möglichkeit zur Überprüfung von Inferenzverfahren dar. Dabei wird ein gentechnisch erzeugtes Teil-Genregulationsnetz in lebende Zellen eingebaut. Werden mit einer solchen Zellkultur gewonnene Genexpressionsdaten zur Inferenz benutzt, kann das bekannte Teil-Genregulationsnetz mit dem zugehörigen Teil des inferierten Netzes verglichen werden.

3.2.1. Künstliche Genregulationsnetze

Mit Hilfe geeigneter Modellsysteme ist es möglich, die Leistungsfähigkeit von Inferenzverfahren zu bewerten. Dafür werden künstliche Genregulationsnetze geschaffen, die ähnliche Eigenschaften wie die natürlichen Systeme aufweisen. Durch Simulation des dynamischen Verhaltens dieser künstlichen Netze werden Messdaten erhalten, die analog den Microarray-Daten zur Inferenz eingesetzt werden. Anschließend kann ein Vergleich (siehe 3.2.3) zwischen inferiertem Netz und künstlichem Genregulationsnetz (Zielnetz) erfolgen, da das Modellsystem bekannt ist (D'HAESELEER, 2000).

Viele Autoren verwenden diese Art der Leistungsbewertung. Die Erzeugung der künstlichen Daten erfolgt auf ebenso vielfältige Art und Weise, wie die oben besprochene Inferenz. So findet man azyklische boolesche Modelle (IDEKER ET AL., 2000), asynchrone (AKUTSU ET AL., 1998b) und synchrone boolesche Netze (LIANG ET AL., 1998; AKUTSU ET AL., 1999) ebenso wie verschiedene Typen von Differenzialgleichungssystemen (z.B. MAKI ET AL., 2001; WEAVER ET AL., 1999; KYODA ET AL., 2000; WAHDE UND HERTZ, 2000; ANDO UND IBA, 2001; MOROHASHI UND KITANO, 1999; D'HAESELEER, 2000; YEUNG ET AL., 2002; WESSELS ET AL., 2001) und Bayes-Netze (Monte Carlo Simulation: IMOTO ET AL., 2002). Die erzeugten künstlichen Messdaten entsprechen der Art der für die Inferenz verwendeten Daten, d.h. es werden Zeitserien, Gleichgewichtsmessungen und diverse Störexperimente simuliert.

Die Erzeugung der künstlichen Netze erfolgt oft **zufällig** (WEAVER ET AL., 1999; IDEKER ET AL., 2000; KYODA ET AL., 2000; LIANG ET AL., 1998; AKUTSU ET AL., 1999; ANDO UND IBA, 2001), d.h. die Kanten und andere Parameter werden zufällig aus einer vorgegebenen Verteilung gezogen. In einigen Fällen werden (kleine) künstliche Netze von Hand **konstruiert** (MAKI ET AL., 2001; IMOTO ET AL., 2002; WAHDE UND HERTZ, 2000; MOROHASHI UND KITANO, 1999) oder bekannte Genregulationsnetze aus der Literatur oder Datenbanken als Simulationsgrundlage genommen (AKUTSU ET AL., 1998b). Einen ähnlichen Weg beschreiten FRIEDMAN ET AL. (2000), die ein inferiertes Netz als Ausgangspunkt für die Erzeugung künstlicher Genexpressionsdaten verwenden.

In vielen Fällen werden zur Bewertung der Methode eine Reihe von zufälligen künstlichen Netzen erzeugt, wobei bestimmte Eigenschaften variiert werden, so z.B. die Größe des Netzes und die Eingangsgradverteilung. Dabei wird häufig durch zusätzliche Forderungen versucht, bestimmte gewünschte Eigenschaften zu sichern. So lassen LIANG ET AL. (1998) und IDEKER ET AL. (2000) nur boolesche Funktionen zu, bei denen eine echte Abhängigkeit des Ausganges von jedem Eingang besteht. WEAVER ET AL. (1999) fordern, daß jedes Gen einen positiven und einen negativen Eingang besitzt. Bei KYODA ET AL. (2000) wird keine Selbstregulation, jedoch indirekte Rückkopplung erlaubt.

Um den Einfluss von in realen Messdaten auftretenden Fehlern auf die Inferenz zu berücksichtigen, fügen einige Autoren den künstlichen Daten Rauschen hinzu. Dieses Rauschen kann normalverteilt (Weaver et al., 1999, jedoch maximal 10% des Messwertes) oder gleichverteilt (Ando und Iba, 2001) sein. Wessels et al., 2001 führen mehrere Versuche mit steigendem Anteil von normalverteiltem Rauschen durch. In manchen Fällen entspricht das zugefügte Rauschen den aus Mehrfachmessungen realer Genexpressionsdaten bekannten Streuungen (D'haeseleer, 2000).

Zum Teil werden neben der Bewertung der Ergebnisse noch andere Ziele mit künstlichen Genexpressionsdaten verfolgt. Insbesondere wird versucht, die für die Inferenz notwendige Anzahl von Messwerten pro Gen abzuschätzen: In Akutsu et al. (1998b) werden so die theoretischen Vorhersagen überprüft, während in Akutsu et al. (1999) mit Hilfe künstlicher Netze die konstanten Faktoren der theoretisch vorhergesagten unteren und oberen Schranken bestimmt werden. Der Einfluss von Rauschen auf die Anzahl benötigter Messwerte wird von Akutsu et al. (2000) mit Hilfe künstlich erzeugter Daten untersucht. Ebenso untersuchen Liang et al. (1998) mit Hilfe von Modellsystemen die benötigte Anzahl von Messwerten zur eindeutigen Netzinferenz. Yeung et al. (2002) ermitteln mit Hilfe künstlicher Genregulationsnetze die Abhängigkeit der notwendigen Anzahl von Messwerten für eine fehlerfreie Netzinferenz von der Größe des Netzwerkes.

3.2.2. Eignung der künstlich erzeugten Testsysteme

Die meisten künstlichen Netze, die zur Bewertung von Inferenzverfahren eingesetzt werden, weichen in zu starkem Maße von natürlich vorkommenden Genregulationsnetzen ab, um eine gute Einschätzung der Leistungsfähigkeit der Algorithmen, angewendet auf reale Daten, zu erhalten. Im Folgenden seien die Hauptpunkte zusammengetragen, in denen die künstlichen Netze von der biologischen Realität zu stark abweichen (vergleiche 2.2):

1. Die Größe der künstlichen Netze ist in vielen Fällen völlig unzureichend, um der Komplexität von biologischen Genregulationsnetzen gerecht zu werden.

   Das betrifft zum einen die Berechnungskomplexität: Verfahren, deren Ziel darin besteht, im Genommaßstab zu inferieren, können mit Netzen, die um mehr als zwei Größenordnungen kleiner als reale Genregulationsnetze sind, kaum hinsichtlich ihrer Effizienz bewertet werden. So sind einige Verfahren (z.B. Morohashi und Kitano, 1999) darauf angewiesen, alle möglichen Netztopologien zu untersuchen. Bei den Experimenten mit ihrem Verfahren verwenden Morohashi und Kitano (1999) Netze mit drei Knoten. Die Untersuchung aller $2^{3^2} = 512$ möglichen Netztopologien ist für diese Größe durchführbar. Da die Anzahl der Netztopologien exponentiell mit der Anzahl der Knoten wächst, ist dies jedoch für größere Netze nicht mehr effizient möglich. In dem Artikel wird auf dieses Problem hingewiesen. Bei größeren Netzen werden, anstelle einer erschöpfenden Suche, zufällig Netztopologien zur weiteren Untersuchung gezogen. Wie auf diese Weise allerdings gesichert wird, dass die gesuchte Netztopologie unter den betrachteten Kandidaten ist, wird nicht beschrieben. Die Überprüfung des Verfahrens anhand kleiner Netze lässt also keine Aussage über das Verhalten bei großen Netzen zu.

   Zum anderen werden künstliche Netze mit deutlich weniger als zehn Knoten (z.B. mit vier Knoten: Wahde und Hertz, 2000) der Komplexität des Problems nicht gerecht. So passen Wahde und Hertz (2000) ihr Modell mit nur 14 Parametern an die

künstlichen Messdaten an. Der Optimierungsalgorithmus müsste bei realen Genregulationsnetzen eine um den Faktor 10^3 höhere Anzahl von Parametern bewältigen.

2. Der oft realisierte geringe Eingangsgrad bei künstlichen Netzen stellt eine starke Reduzierung der Komplexität dar. Manche Autoren verwenden sogar Netze mit konstantem Eingangsgrad (Akutsu et al., 1999, Eingangsgrad zwei bzw. drei), was zu einer weiteren Verminderung der Komplexität führt. Dadurch wird die bei ihrem Verfahren notwendige vollständige Suche im Raum der booleschen Funktionen erst effizient möglich. Diese sehr geringe Konnektivität ist biologisch unplausibel.

3. Einige der künstlichen Netze sind zyklenfrei oder weisen sehr wenige Zyklen auf (Maki et al., 2001; Ideker et al., 2000; Imoto et al., 2002). Dies steht im starken Gegensatz zur biologischen Realität. Zyklische und azyklische Netze können völlig verschiedenes dynamisches Verhalten aufweisen. Ein Test eines Verfahrens für die Inferenz stark zyklischer Netze (wie den Genregulationsnetzen) mit azyklischen Netzen ist somit höchst inadäquat. Bei Kyoda et al. (2000) werden künstliche Netze ohne Selbstregulation erzeugt. Dies stellt zwar eine geringere Einschränkung als Zyklenfreiheit dar, blendet aber einen biologisch häufig anzutreffenden Mechanismus völlig aus.

4. Die Erzeugung von künstlichen Daten auf der Grundlage von booleschen Netzen (Liang et al., 1998; Ideker et al., 2000) stellt eine zu starke Vereinfachung der Vorgänge bei der Genexpression dar.

5. Posttranskriptionale Regulation und Translation werden meist nicht berücksichtigt. Die Aktivitäten eines Primärtranskripts (welche im Microarray gemessen wird) und des zugehörigen Proteins (welches auf die Expression anderer Gene einen Einfluss ausübt) können jedoch sehr verschieden sein.

6. Bei der Genexpression spielen oft Dimere oder komplexere Verbindungen eine regulierende Rolle. Dem Einfluss eines Gens auf ein anderes ist somit eine weitere Stufe zwischengeschaltet. Diese wird in keinem der künstlichen Genregulationsnetze betrachtet.

Ein weiterer Kritikpunkt besteht darin, dass die Erzeugung von künstlichen Daten oft mit derselben Art von Modell erfolgt, auf dem auch die Inferenz beruht. Eine gute Leistung auf den so gewonnenen künstlichen Daten spricht nur dafür, dass das Verfahren in der Lage ist, diese Art von Daten zu repräsentieren. Weder die Eignung der Art von Modellierung zur Repräsentation von biologischen Genregulationsnetzen noch die Fähigkeit des Inferenzverfahrens auf realen Genexpressionsdaten kann so gezeigt werden: Selbst ein Inferenzverfahren, das auf einem biologisch völlig unplausiblen Modell der Genexpression basiert, könnte bei von diesem Modell erzeugten künstlichen Daten hervorragende Ergebnisse erzielen.

Die mit konstruierten (nicht zufällig erzeugten) künstlichen Genregulationsnetzen (Maki et al., 2001; Wahde und Hertz, 2000; Wessels et al., 2001) gewonnenen Testdaten lassen nur begrenzte Aussagen über die Leistungsfähigkeit der untersuchten Methoden zu. Anstatt an einem solchen Spezialfall zu testen, sollte mit Hilfe vieler zufällig erzeugter Netze eine Testreihe durchgeführt werden. Auf diese Weise sind statistische Aussagen über die Wirksamkeit des Verfahrens auf neuen, ähnlichen Daten möglich.

Obwohl, wie eben beschrieben, die meisten künstlichen Netze nicht geeignet zur Bewertung der Inferenzverfahren sind, gibt es doch einige positive Ansätze:

D'haeseleer (2000) erzeugt zwar relativ kleine Netze (20 Knoten mit mittlerem Eingangsgrad vier), orientiert sich dabei jedoch an vorliegenden realen Messdaten. Dabei wird nicht nur ein den realen Messfehlern ähnliches Rauschen den künstlichen Daten zugefügt, sondern es wird auch versucht, einige globale Eigenschaften der realen Messdaten künstlich nachzuempfinden.

Die meines Erachtens gründlichste Untersuchung ihres Verfahren führen Yeung et al. (2002) durch. Sie verwenden verschiedene Arten künstlicher Testsysteme:

1. Lineare Netzwerke: Die Eingangsgrade werden entsprechend einer beschränkten Potenzgesetzverteilung gewählt (Jeong et al., 2000), wobei der Maximalwert viel kleiner als die Netzgröße ist. Die Simulation des Netzwerkes erfolgt mit dem entsprechenden linearen Differenzialgleichungssystem.

2. Repressionskaskade: Eine Sequenz von Genen, wobei jedes Gen das folgende in seiner Expression hemmt. Diese wird durch eine entsprechende Reaktionskinetik (nichtlineares Differenzialgleichungssystem) beschrieben.

3. Künstliches Genregulationsnetz: Die Eingangsgrade werden wie beim linearen Netzwerk gewählt, die Simulation erfolgt hier jedoch mit einer nichtlinearen Reaktionskinetik.

Entsprechend diesen drei Modellen wurden künstliche Systeme mit 100 bis 1000 Genen erzeugt, wobei die Parameter und Eingangsgrade zufällig gewählt werden.

3.2.3. Bewertungsmaße

Hier werden Maße für die Bewertung von Inferenzmethoden angeführt. Insbesondere wird dabei auf den Vergleich zwischen inferiertem Netz und (künstlichem) Zielnetz eingegangen. Die Daten für die Inferenz werden vom künstlichen Zielnetz erzeugt.

Zuerst müssen einige Begriffe für Kanten eingeführt werden: Dafür stelle man sich ein Genregulationsnetz mit n Genen als Matrix $W_{n\times n}$ vor. Bei booleschen Netzmodellen ist der Eintrag $W_{i,j}$ $(0 < i, j \leq n)$ gleich eins, falls eine Kante von i nach j existiert, anderenfalls Null.

Bei kontinuierlichen Modellen stehen in der Matrix die Kantengewichte. Eine Kante (i, j) heißt **vorhanden**, wenn $|W_{i,j}|$ über einem festgelegten Schwellwert liegt, anderenfalls **nicht vorhanden**. Eine (nicht) vorhandene Kante in einem inferierten Netz heißt **(nicht) gefunden**.

Richtig inferierte Kanten (true positive, TP**)** Anzahl der gefundenen Kanten, die auch im Zielnetz vorhanden sind (Weaver et al., 1999).

Falsch inferierte Kanten (false positive, FP**)** Anzahl der gefundenen Kanten, die im Zielnetz nicht vorhanden sind (Weaver et al., 1999).

Richtig nicht inferierte Kanten (true negative, TN**)** Anzahl der nicht gefundenen Kanten, die auch im Zielnetz nicht vorhanden sind (Ando und Iba, 2001).

Falsch nicht inferierte Kanten (false negative, FN**)** Anzahl der nicht gefundenen Kanten, die im Zielnetz vorhanden sind (Ando und Iba, 2001).

Damit ist eine saubere Definition der beiden folgenden in der Literatur oft verwendeten Maße möglich:

Die **Sensitivität** (SN) ist ein Maß für die Vollständigkeit der Lösung. Nach Ando und Iba (2001) und Kyoda et al. (2000) ist sie der Anteil der vorhandenen Kanten im Zielnetz, die auch gefunden wurden:

$$\mathrm{SN} =_{df} \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad . \tag{3.4}$$

Die **Spezifität** (SP) ist ein Maß für die Korrektheit der Lösung. Während Kyoda et al. (2000) sie als Anteil der gefundenen Kanten an den vorhandenen Kanten im Zielnetz definieren:

$$\mathrm{SP} =_{df} \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{3.5}$$

ist sie bei Ando und Iba (2001) der Anteil der richtig nicht inferierten Kanten an den nicht vorhandenen Kanten im Zielnetz:

$$\mathrm{SP} =_{df} \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}} \quad . \tag{3.6}$$

Yeung et al. (2002) nennen eine Kante (i, j) im inferierten Netz **korrekt**, falls $|W_{i,j} - \widehat{W}_{i,j}| \leq \delta$, wobei $\delta > 0$ eine festgelegte Fehlertoleranz darstellt, W die Matrix des Zielnetzes und $\widehat{W}$ die Matrix des inferierten Netzes ist.

Wessels et al. (2001) vergleichen mehrere Inferenzmethoden, die auf kontinuierlichen Modellen beruhen. Sie verwenden dafür mehrere Arten von Differenzialgleichungen als Grundlage für die Erzeugung von künstlichen Daten: ein lineares Modell (analog Yeung et al., 2002), ein sigmoides Modell entsprechend Wahde und Hertz (2000) und das nichtlineare Modell von Savageau (1998), welches sie als am meisten geeignet betrachten. Mit Hilfe dieser Modellen werden Zeitserien simuliert. Dafür wird die Verbindungsmatrix eines Netzes mit fünf Knoten vorgegeben und die Startzustände zufällig gewählt. Den künstlichen Messdaten werden anschließend verschiedene Stufen von Rauschen hinzugefügt. Die Bewertung erfolgt mit Hilfe folgender Maße (für Details siehe Wessels et al., 2001, S. 5ff):

Schlußvermögen (*inferential Power*) ist ein Maß für die Ähnlichkeit zwischen Ziel- und inferierter Gewichtsmatrix und entspricht dem Korrektheitsbegriff von Kanten von Yeung et al. (2002) (siehe oben).

Vorhersagefähigkeit (*predictive Power*) ist ein Maß für die Nähe zwischen den Ausgangsmessdaten (vom Zielnetz erzeugt) und den entsprechenden vom inferierten Netz erzeugten Daten.

Robustheit ist eine Bewertung der Anfälligkeit des Ergebnisses auf Störungen der Messdaten.

Konsistenz ist ein Maß für die Ähnlichkeit aller inferierbarer Netze, wenn ein Messdatensatz gegeben ist, der zur eindeutigen Netzinferenz unzureichend ist.

Stabilität beschreibt die Fähigkeit eines Netzmodells, die Zustandsvariablen (Konzentrationen) innerhalb vernünftiger Grenzen zu halten.

Berechnungskosten werden bei Wessels et al. (2001) einfach durch Messung der Berechnungszeit für die Lösung konkreter Problemen bestimmt.

4. Künstliche Genexpressionsdaten

Wie in den vorangegangenen Abschnitten gezeigt wurde, existieren eine Fülle von Inferenzverfahren für Genregulationsnetze und verschiedene Ansätze, diese zu bewerten. Aufgrund mangelndem Wissens über biologische Genregulationsnetze wird eine Überprüfung häufig anhand von Modellsystemen (künstliche Genregulationsnetze) vorgenommen. Die meisten der zur Bewertung benutzten künstlichen Systeme scheinen jedoch für diesen Zweck unzureichend, da sie die biologische Realität nur dürftig wiedergeben (siehe 3.2.2).

In diesem Kapitel wird ein geeigneteres Modellsystem entwickelt. Es soll in der Lage sein, künstliche Netze zu erzeugen, die in ihrer Komplexität an biologische Genregulationsnetze heranreichen. Durch Simulation dieser künstlichen Netze können Testdaten bereitgestellt werden, die eine (vergleichende) Bewertung von Inferenzverfahren ermöglichen. Bestimmte Eigenschaften der künstlichen Genregulationsnetze und der daraus erzeugten Daten können durch entsprechende Parameter eingestellt werden.

Die Modellierung (Abschnitt 4.1) orientiert sich an den Vorgängen bei der Genexpression. Insbesondere werden sowohl die Transkription als auch die Translation und die Regulation auf diesen beiden Ebenen berücksichtigt. Jede Substanz unterliegt außerdem einem Zerfallsprozess. Eine Zuordnung zwischen einer strukturellen Darstellung als künstliches Genregulationsnetz und einer kinetischen Beschreibung mittels eines Differenzialgleichungssystems wird ermöglicht. Entsprechend dieser Trennung zwischen Struktur und Dynamik kann auch die Erzeugung der künstlichen Genexpressionsdaten erfolgen (Abschnitt 4.2): Im ersten Schritt wird ein geeignetes künstliches Genregulationsnetz erzeugt und anschließend sein dynamisches Verhalten durch Lösen eines entsprechenden Differenzialgleichungssystems simuliert. Der zweite Schritt ermöglicht durch Wahl der Anfangswerte und Modifikationen des Differenzialgleichungssystems die Simulation beliebiger experimenteller Messungen inklusive Störexperimenten. Die Umsetzung in ein `Java`-Programm wird im letzten Abschnitt besprochen (4.3).

Obwohl das hier vorgestellte Modell in einigen Punkten von der üblichen Modellierung von Genregulationsnetzen abweicht, stellt es keine Einschränkung dar: Mit Definition 4.12 wird gezeigt, dass die hier gewählte Darstellung auf eine klassische Darstellung eines Genregulationsnetzes als gerichteter Graph reduziert werden kann. Somit kann das künstliche Genregulationsnetz in der üblichen Form ausgegeben und mit einem inferiertem Netz verglichen werden.

4.1. Modellierung der Genexpression

In diesem Abschnitt wird ein abstraktes mathematisches Modell der Genexpression vorgestellt. Die biologischen Vorgänge werden in vereinfachter Weise durch klar definierte Begriffe beschrieben. Das erleichtert die Diskussion über diese Vorgänge und erlaubt die Ableitung mathematischer Aussagen über das Modell.

Die Regulation der beiden Vorgänge Transkription und Translation wird durch eine identische Struktur, eine „Produktionseinheit“, beschrieben. Jedem Gen und jeder kodierenden RNS ist eine solche Produktionseinheit zugeordnet. Aus einer Menge solcher Produktionseinheiten wird der zentrale Begriff des künstlichen Genregulationssystems abgeleitet. Die Menge der Produktionseinheiten enthält nur Informationen über die strukturellen Eigenschaften des Systems: Welches Gen reguliert mittels welcher Substanz die Expression welchen anderen Gens? Dies ermöglicht die Abbildung auf die klassische Darstellung eines

Genregulationsnetzes. Dagegen erfolgt durch die Definition als Genregulationssystem eine Übertragung in eine Menge miteinander gekoppelter Differenzialgleichungen und somit in eine Beschreibung der Dynamik des Systems.

Die Grundzüge und Annahmen der Modellierung lassen sich wie folgt zusammenfassen, die *hervorgehobenen* Begriffe werden im Anschluss formal definiert:

1. In der Zelle gibt es eine Vielzahl von biologisch aktiven *Substanzen*. Es wird angenommen, dass jede Substanz in einer bestimmten *Konzentration* vorliegt, die sich im Laufe der Zeit ändern kann. Die meisten dieser Substanzen unterliegen einem Zerfallsprozess. Im folgenden Modell werden *einfache* (*Proteine* und *RNS*) und *zusammengesetzte* Substanzen (Dimere und kompliziertere Gebilde) betrachtet. Dabei wird davon ausgegangen, dass eine RNS, insofern sie translatiert wird, ein eindeutig bestimmtes Protein *produziert*.

2. Der *Zerfall* einer Substanz wird als Reaktion 1. Ordnung betrachtet, wobei die Zerfallsprodukte unberücksichtigt bleiben.

3. Für jede Substanz existiert eine zeitlich variable *externe Änderung* ihrer Konzentration.

4. Die DNS, die ebenfalls in der Zelle vorhanden ist, wird im folgenden Modell nicht explizit betrachtet. Vielmehr werden die durch sie hervorgerufenen Prozesse modelliert.

5. Ein *Gen* wird in Abhängigkeit der Konzentrationen bestimmter Substanzen (*Transkriptionsregulatoren*) in die zugehörige RNS transkribiert. Die damit verbundene Konzentrationsänderung dieser RNS ist eine Funktion (*Transkriptionsfunktion*) der Konzentrationen der regulierenden Substanzen.

6. In Abhängigkeit von den Konzentrationen bestimmter Substanzen (*posttranskriptionale Regulatoren*) erfolgt die Translation einer RNS in das zugehörige Protein. Die damit verbundene Konzentrationsänderung dieses Proteins ist wiederum eine Funktion (*Translationsfunktion*) der Konzentrationen der regulierenden Substanzen und der Konzentration der zugehörigen RNS.

7. Die Abnahme der Konzentration einer regulatorischen Substanz durch den Regulationsvorgang (z.B. Anlagerung an ein geeignetes cis-Element) wird vernachlässigt.

8. Sowohl die Transkription eines Gens in eine RNS als auch die Translation dieser RNS in ein Protein, werden jeweils durch eine *Produktionseinheit* beschrieben. Somit ist jedem Gen eine Transkriptionseinheit und somit eine RNS (Primärtranskript) zugeordnet. Zu jeder kodierenden RNS gehört eine Translationseinheit und somit ein produziertes Protein.

9. Das *Genom*, als eine Menge von Genen, impliziert ein System von sich gegenseitig beeinflussenden Prozessen. Aus diesem wird ein Differenzialgleichungssystem (*Genregulationssystem*) abgeleitet: Für jede Substanz existiert eine Differenzialgleichung, die die zeitliche Änderungsrate ihrer Konzentration angibt.

10. Es wird angenommen, dass sich die zeitliche Änderungsrate der Konzentration einer Substanz additiv aus den entsprechenden Änderungsraten aller sie erzeugenden oder abbauenden Prozesse zusammensetzt.

4.1.1. Komponenten und Zustand

Der Zustand einer Zelle kann durch die Konzentrationen aller in ihr potenziell vorhandenen Substanzen charakterisiert werden. Dafür muss die Menge dieser Substanzen festgelegt werden.

Zunächst erfolgt eine Definition der Menge der einfachen Substanzen: RNS und Proteine. Zwischen den RNS und den Proteinen gibt es eine eindeutige Zuordnung: Jede RNS produziert ein bestimmtes Protein[1]. Es gibt in der Zelle jedoch auch RNS, die nicht für ein Protein kodiert, sondern eine andere Funktion übernimmt. Die Zuordnung müsste aus diesem Grund partiell sein. Um die Produktfunktion total zu machen, wird eine spezielles Objekt $\perp$ eingeführt, das keine Substanz darstellt. Alle nicht kodierenden RNS „produzieren“ dieses Objekt.

Definition 4.1 (Einfache Substanzen)

Gegeben seien zwei disjunkte, endliche, nichtleere Mengen $\mathrm{S_{RNS}}$ *und* $\mathrm{S_{Prot}}$*, deren Elemente als* **RNS** *bzw.* **Proteine** *bezeichnet werden. Die Vereinigung der beiden Mengen heißt Menge der* **einfachen Substanzen***:* $\mathrm{S}_0 =_{df} \mathrm{S_{RNS}} \cup \mathrm{S_{Prot}}$*. Zwischen den beiden Mengen gebe es die* **Produktfunktion** $P : \mathrm{S_{RNS}} \to \mathrm{S_{Prot}} \cup \{\perp\}$*, die einer RNS das von ihr* **produzierte** *Protein bzw.* $\perp$ *zuordnet. Gilt* $P(m) = \perp$ *für* $m \in \mathrm{S_{RNS}}$ *heißt* m **nicht kodierend** *anderenfalls* **kodierend***.*

Aus diesen einfachen Substanzen können zusammengesetzte Substanzen gebildet werden. Für zwei einfache Substanzen beschreibt eine Funktion $\gamma_0 : \mathrm{S}_0^2 \to [0, \infty)$ die Neigung, ein Dimer zu bilden. Die Menge der Dimere ist die Menge all jener Paare von einfachen Substanzen, die eine nicht verschwindende Neigung zur Dimerisation haben. Aus diesen Dimeren und den einfachen Substanzen können nun analog Trimere und daraus immer kompliziertere Gebilde zusammengesetzt werden.

Vereinfachend wird hier angenommen, dass zwei Substanzen maximal eine zusammengesetzte Substanz bilden können. Aus diesem Grund erfolgt die mathematische Repräsentation einer zusammengesetzten Substanz als Multimenge. In einer Multimenge spielt (wie in einer Menge) die Reihenfolge der Elemente keine Rolle, d.h. die Multimengen $\langle x, y\rangle$ und $\langle y, x\rangle$ sind gleich. Eine Multimenge ermöglicht, im Gegensatz zur Menge, das mehrfache Auftreten von Elementen. So kann ein Homodimer aus der Substanz s als Multimenge $\langle s, s\rangle$ dargestellt werden. Da die Funktionen γ_i für Paare von Substanzen definiert sind, wird zusätzlich Symmetrie gefordert: Für zwei Substanzen $s, q \in \mathrm{S}_i$ gilt damit $\gamma_i(s, q) = \gamma_i(q, s)$. Ist $\gamma_i(s, q) > 0$, existiert genau eine zusammengesetzte Substanz $\langle s, q\rangle \in \mathrm{S}_{i+1}$.

Definition 4.2 (Multimenge)

Gegeben sei eine beliebige Menge X*. Eine* **Multimenge** M_X *ist eine ungeordnete Anordnung von Elementen aus* X*, wobei ein Element auch mehrmals auftreten darf. Formal kann sie als Funktion* $M_X : X \to \mathbb{N}$ *beschrieben werden, wobei der Wert* $M_X(x)$ *angibt, wie oft* $x \in X$ *in* M_X *vertreten ist.*

Eine zweielementige Multimengen über einer Menge X *wird mit* $\langle x, y\rangle$ *für* $x, y \in X$ *notiert.*

[1] Munk (2001, S. 2-36) weist darauf hin, dass durch einen nach der Transkription stattfindenden Vorgang („alternatives Spleißen“), die Zuordnung „ein Gen entspricht einem Protein“ nicht gelten muss: Ein Gen kann für mehrere Proteine kodieren. Durch Einführung einer Funktion $\mathrm{aSpl} : \mathrm{S_{RNS}} \to \mathcal{P}(\mathrm{S_{RNS}})$ und Anpassung der Definitionen 4.16 und 4.17 kann die Modellierung so erweitert werden, dass das alternative Spleißen berücksichtigt wird. Darauf wird hier jedoch verzichtet.

Definition 4.3 (Substanzen)

Die Menge S *der* **Substanzen** *wird wie folgt induktiv definiert:*

(Induktionsanfang) $\mathrm{S}_0 =_{df} \mathrm{S}_{\mathrm{RNS}} \cup \mathrm{S}_{\mathrm{Prot}}$.

(Induktionsschritt) *Für $i \geq 0$ sei eine symmetrische Funktion $\gamma_i : \mathrm{S}_i^2 \to [0, \infty)$ für Paare aus S_i gegeben, die die Neigung zur Zusammenlagerung der beiden Substanzen beschreibt. Die Menge S_{i+1} wird wie folgt definiert:*

$$\mathrm{S}_{i+1} \quad =_{df} \quad \mathrm{S}_i \cup \{\langle s, q\rangle : s, q \in \mathrm{S}_i \wedge \gamma_i(s, q) > 0\}. \tag{4.1}$$

(Induktionsschluss) *Die Menge* S *aller Substanzen ist die kleinste obere Schranke aller dieser Mengen:*

$$\mathrm{S} \quad =_{df} \quad \lim_{i \to \infty} \mathrm{S}_i \,. \tag{4.2}$$

Jede Substanz, die nicht einfach ist, heißt **zusammengesetzt**.

Mit dieser Definition sind sehr komplexe Strukturen „konstruierbar“, z.B. beliebig lange Polymere. Die Menge S kann je nach Wahl der γ_i beliebig groß werden, sogar unendlich.

Jede Substanz liegt zu einem bestimmten Zeitpunkt in einer bestimmten Quantität (Konzentration) im System vor:

Definition 4.4 (Konzentration)

Zu jeder Substanz $s \in \mathrm{S}$ wird durch $X : \mathrm{S} \to (\mathbb{R} \to [0, \infty))$ eine Funktion bestimmt: $x_s : \mathbb{R} \to [0, \infty)$ ordnet jeder reellen Zahl $t \in \mathbb{R}$ eine positive reelle Zahl $x_s(t)$, die **Konzentration** *von s zum* **Zeitpunkt** *t, zu.*

Neben den Substanzen enthält das System noch weitere Komponenten, die Gene:

Definition 4.5 (Gen, Genom)

Gegeben sei eine endliche, nichtleere Menge $\mathfrak{G}$, deren Elemente als **Gene** *bezeichnet werden. Jedem Gen $G \in \mathfrak{G}$ wird ein* **Primärtranskript** *$m \in \mathrm{S}_{\mathrm{RNS}}$ zugeordnet. $\mathfrak{G}$ heißt auch* **Genom**.

4.1.2. Prozesse

Nach der Definition der Komponenten (Genom und Substanzen) und der Beschreibung des Zustands des Systems über die Konzentrationen aller Substanzen werden nun die Prozesse festgelegt, die den Zustand des Systems ändern können. Eine Übersicht über alle hier betrachteten Prozesse gibt Abbildung 4.6 auf Seite 50.

Für jede Substanz gibt es zwei elementare Prozesse: den Zerfall und die externe Änderung. Während der Zerfall durch einen konstanten Wert charakterisiert wird, kann die externe Änderung zeitlich variieren.

Definition 4.6 (Elementare Prozesse)

Zu jeder Substanz $s \in \mathrm{S}$ gibt es zwei elementare Prozesse, die ihre Konzentration ändern können. Diese sind folgendermaßen bestimmt:

(6–1) *Die Rate des* **Zerfalls** *wird durch eine Konstante $\delta_s \geq 0$ charakterisiert.*

(6–2) *Durch eine Funktion $y_s : \mathbb{R} \to \mathbb{R}$ wird die* **externe Änderung** *von s zum Zeitpunkt t beschrieben.*

Bei der oben beschriebenen Zusammenlagerung von Substanzen finden ebenfalls Prozesse statt: die Bildung der zusammengesetzten Substanz und der Verbrauch der Ausgangssubstanzen.

Definition 4.7 (Prozesse bei der Zusammenlagerung)

Falls für zwei Substanzen $s, q \in \mathrm{S}_i$ gilt $\gamma_i(s, q) > 0$, so existieren folgende Prozesse:

(7–1) *Die* **Bildung** *von $\langle s, q \rangle$ aus s und q.*

(7–2) *Der* **Verbrauch** *von s und q* **durch Bildung** *von $\langle s, q \rangle$.*

Einen weiteren komplexen Prozess stellt die Expression eines Gens, d.h. die Übersetzung der DNS in ein Protein, dar. Dieser Prozess besteht aus zwei Schritten: Transkription und Translation. Beide Schritte können durch das Vorhandensein von bestimmten Substanzen beeinflusst werden.

In einer Produktionseinheit werden die Informationen für einen solchen Teilprozess zusammengefasst: Es wird festgelegt, welche RNS bzw. welches Protein produziert wird und welche Substanzen diesen Vorgang beeinflussen. Die Produktionsfunktion beschreibt die Änderungsrate der Konzentration des Produktes in Abhängigkeit der Konzentrationen der Regulatoren (siehe auch Definition 4.16) und (bei Translation) der entsprechenden RNS.

Definition 4.8 (Produktionseinheit)

Eine **Produktionseinheit** Π *ist eine Struktur* $\Pi = (s, \mathrm{R}, \mathrm{Prd})$ *mit folgenden Eigenschaften:*

(8–1) $s \in \mathrm{S}_{\mathrm{RNS}} \cup \mathrm{S}_{\mathrm{Prot}}$ *ist das* **Produkt** *von* Π.

(8–2) $\mathrm{R} = \{s_1, \ldots, s_l\} \subseteq \mathrm{S}$ *ist die (endliche) Menge der* **Regulatoren** *von* Π.

(8–3) Prd *ist eine l- bzw. $(l+1)$-stellige Funktion, die* **Produktionsfunktion***:*

$$\mathrm{Prd} \quad : \quad \begin{cases} (\mathbb{R}^+)^l \to \mathbb{R} & \text{falls } s \in \mathrm{S}_{\mathrm{RNS}}, \\ (\mathbb{R}^+)^{l+1} \to \mathbb{R} & \text{falls } s \in \mathrm{S}_{\mathrm{Prot}}. \end{cases} \tag{4.3}$$

Es gibt zwei Typen von Produktionseinheiten, die sich in der Sorte des Produktes unterscheiden: Transkriptionseinheiten produzieren RNS, Translationseinheiten Proteine.

Definition 4.9 (Transkription, Translation)

(9–1) *Jedem Gen* $G \in \mathfrak{G}$ *mit Primärtranskript* $m \in \mathrm{S}_{\mathrm{RNS}}$ *sei eine Produktionseinheit* $\Pi_G = (m, \mathrm{R}_G, \mathrm{Prd}_G)$ *zugeordnet, die als* **Transkriptionseinheit** *von* G *bezeichnet wird.* R_G *heißt Menge der* **Transkriptionsregulatoren** *von* Π_G*.*

(9–2) *Jeder kodierenden RNS* $m \in \mathrm{S}_{\mathrm{RNS}}$ *mit Produkt* $P(m) \in \mathrm{S}_{\mathrm{Prot}}$ *sei eine Produktionseinheit* $\Pi_m = (P(m), \mathrm{R}_m, \mathrm{Prd}_m)$ *zugeordnet, die als* **Translationseinheit** *von* m *bezeichnet wird.* R_m *heißt Menge der* **posttranskriptionalen** *Regulatoren von* Π_m*.*

4.1.3. Klassische Beschreibung

Einem Genom ist somit eine Menge von Transkriptionseinheiten zugeordnet, die die Übertragung der Information der Gene in RNS beschreibt. Für die kodierenden RNS charakterisieren die Translationseinheiten die sich anschließende Übersetzung in Proteine. Somit ist für jedes Gen festgelegt, für welche RNS und damit für welches Protein es kodiert und wie die Regulation der beiden Teilschritte erfolgt.

Diese Modellierung weicht von der herkömmlichen Darstellung eines Genregulationsnetzes ab, da die Kanten nur implizit vorhanden sind. Im folgenden wird beschrieben, wie die einem Genom zugehörige Menge von Produktionseinheiten in die übliche Darstellung umgewandelt werden kann. Diese Darstellung wird hier als GRN-Graph bezeichnet. Ein GRN-Graph ist eine Menge von Genen (Knoten) und direkten Einflüssen zwischen diesen, die durch Kanten repräsentiert werden. Die Modellierung erlaubt multiple Regulationswege zwischen zwei Genen: Die Regulation kann durch verschiedene produzierte Substanzen vermittelt werden oder beide Prozesse, Transkription und Translation, betreffen. Im multiplen GRN-Graph wird jeder Einfluss durch eine Kante dargestellt. Im einfachen GRN-Graph existiert maximal eine Kante zwischen zwei Genen.

Ein direkter Einfluss auf ein Gen erfolgt über die Regulatoren der Transkription und (eventuell) der Translation. Alle Substanzen, die einen direkten Einfluss auf ein Gen ausüben, werden als direkte Regulatoren des Gens bezeichnet:

Definition 4.10 (Direkte Regulatoren eines Gens)

Gegeben sei ein Gen $G \in \mathfrak{G}$ *mit zugehörigem Primärtranskript* $m \in \mathrm{S}_{\mathrm{RNS}}$ *und Transkriptionseinheit* $\Pi_G = (m, \mathrm{R}_G, \mathrm{Prd}_G)$*. Die Menge* $\mathrm{R}(G) \subseteq \mathrm{S}$ *der* **Regulatoren** *(transkriptional und posttranskriptional) von* G *ist definiert als:*

$$\mathrm{R}(G) \quad =_{df} \quad \begin{cases} \mathrm{R}_G \cup \mathrm{R}_m & \textit{falls } P(m) \neq \perp, \\ \mathrm{R}_G & \textit{sonst}, \end{cases} \tag{4.4}$$

wobei für kodierendes m *die Translationseinheit* $\Pi_m = (P(m), \mathrm{R}_m, \mathrm{Prd}_m)$ *sei.*

Bestimmt man für jeden Regulator eines Gens, diejenigen Gene, die diesen Regulator produzieren, kann man die direkten Einflüsse ermitteln. Eine einfache Substanz wird durch alle Produktionseinheiten produziert, die diese Substanz als Produkt besitzen. Die Produzenten einer zusammengesetzten Substanz sind die Produzenten ihrer beiden Bestandteile. Dies führt zu folgender induktiven Definition:

Definition 4.11 (Produzenten einer Substanz)

Für jede Substanz $s \in \mathrm{S}$ wird die Menge $\mathfrak{G}_{\mathrm{Prod}}(s) \subseteq \mathfrak{G}$ der Gene, die diese Substanz produzieren, wie folgt definiert:

(11–1) *Für jede RNS $m \in \mathrm{S}_{\mathrm{RNS}}$ sei die Menge $\mathfrak{G}_{\mathrm{Prod}}(m) \subseteq \mathfrak{G}$ der* **Produzenten** *von m definiert als alle Gene, die m als Primärtranskript besitzen:*

$$\mathfrak{G}_{\mathrm{Prod}}(m) \quad =_{df} \quad \{G \in \mathfrak{G} : \Pi_G = (m, \mathrm{R}_G, \mathrm{Prd}_G)\}. \tag{4.5}$$

(11–2) *Für jedes Protein $p \in \mathrm{S}_{\mathrm{Prot}}$ sei die Menge $\mathfrak{G}_{\mathrm{Prod}}(p) \subseteq \mathfrak{G}$ der* **Produzenten** *von p definiert als alle Gene, deren Primärtranskript p als Produkt besitzen:*

$$\mathfrak{G}_{\mathrm{Prod}}(p) \quad =_{df} \quad \{G \in \mathfrak{G} : \Pi_G = (m, \mathrm{R}_G, \mathrm{Prd}_G) \wedge P(m) = p\}. \tag{4.6}$$

(11–3) *Für jede zusammengesetzte Substanz $\langle s, q\rangle \in \mathrm{S} \setminus \mathrm{S}_0$ sei die Menge $\mathfrak{G}_{\mathrm{Prod}}(\langle s, q\rangle) \subseteq \mathfrak{G}$ der* **Produzenten** *von $\langle s, q\rangle$ induktiv definiert als:*

$$\mathfrak{G}_{\mathrm{Prod}}(\langle s, q\rangle) \quad =_{df} \quad \mathfrak{G}_{\mathrm{Prod}}(s) \cup \mathfrak{G}_{\mathrm{Prod}}(q). \tag{4.7}$$

Jeder Produzent einer Substanz übt auf alle Gene einen direkten Einfluss aus, die diese Substanz als Regulator besitzen. Dies wird im GRN-Graph als gerichtete Kante dargestellt:

Definition 4.12 (Induzierter (multipler) GRN-Graph)

Gegeben sei ein Genom $\mathfrak{G} = \{G_1, \ldots, G_N\}$ von N Genen.

(12–1) *Der von $\mathfrak{G}$* **induzierte einfache GRN-Graph** *(V, E) wird wie folgt definiert:*

1. *Die Menge der Knoten V wird mit der Menge der Gene $\mathfrak{G}$ identifiziert:*

$$V \quad \equiv \quad \mathfrak{G}. \tag{4.8}$$

2. *Die Menge der Kanten E wird über folgende Äquivalenz bestimmt: Zwischen zwei Knoten $G_1, G_2 \in V$ mit $\Pi_{G_1} = (m_1, \mathrm{R}_{G_1}, \mathrm{Prd}_{G_1})$ und $\Pi_{G_2} = (m_2, \mathrm{R}_{G_2}, \mathrm{Prd}_{G_2})$ gibt es eine Kante $(G_1, G_2) \in E$ genau dann, wenn G_1 Produzent eines direkten Regulators von G_2 ist. Formal:*

$$(G_1, G_2) \in E \quad \longleftrightarrow \quad \exists s \in \mathrm{R}(G_2) : G_1 \in \mathfrak{G}_{\mathrm{Prod}}(s). \tag{4.9}$$

(12–2) *Der von $\mathfrak{G}$* **induzierte multiple GRN-Graph** *(V, E) wird wie folgt definiert:*

1. *Die Menge der Knoten V wird mit der Menge der Gene $\mathfrak{G}$ identifiziert:*

$$V \quad \equiv \quad \mathfrak{G}. \tag{4.10}$$

2. *Die Multimenge der Kanten E wird wie folgt bestimmt: Zwischen zwei Knoten $G_1, G_2 \in V$ mit $\Pi_{G_1} = (m_1, \mathrm{R}_{G_1}, \mathrm{Prd}_{G_1})$ und $\Pi_{G_2} = (m_2, \mathrm{R}_{G_2}, \mathrm{Prd}_{G_2})$ gibt es für jeden direkten Regulator von G_2, von dem G_1 Produzent ist, eine Kante $(G_1, G_2) \in E$. Formal kann dies über die charakteristische Funktion M_E von E beschrieben werden:*

$$\begin{aligned} M_E((G_1, G_2)) =_{df} \; & card\big(\{s \in R_{G_2} : G_1 \in \mathfrak{G}_{\mathrm{Prod}}(s)\}\big) + \\ & card\big(\{s \in R_{m_2} : G_1 \in \mathfrak{G}_{\mathrm{Prod}}(s)\}\big), \end{aligned} \tag{4.11}$$

wobei für kodierendes m_2 die Translationseinheit $\Pi_{m_2} = (P(m_2), \mathrm{R}_{m_2}, \mathrm{Prd}_{m_2})$ sei und für nicht kodierendes m_2 gelte $\mathrm{R}_{m_2} =_{df} \emptyset$.

Wird für die Produktionsfunktion in Definition 4.8 verlangt, dass sie in jeder Stelle monoton ist, ist eine Einteilung der Kanten im multiplen GRN-Graph in aktivierende und inhibierende möglich. Im einfachen Graph, wo jede Kante mehrere regulierende Einflüsse repräsentieren kann, ist dies nicht mehr durchführbar.

Durch diese Trennung der Genexpression in zwei unabhängige Prozesse (Transkription und Translation) kann es zu einer gemeinsamen Nutzung von Regulatoren durch Gene kommen (siehe auch Beispiel 1):

1. Gene mit gleichem Primärtranskript (im Beispiel die Gene A und B) besitzen dieselbe Translationseinheit und somit identische posttranskriptionale Regulatoren. Damit sind sie bei der Translation durch die gleichen Gene beeinflusst (im Beispiel durch Gen C).

2. Ebenso besitzen Gene mit gleichem Primärtranskript identische Mengen von Produkten und regulieren somit dieselben Produktionseinheiten (im Beispiel wird Gen D von den Genen A und B gemeinsam reguliert).

Gene mit gleichem Primärtranskript sind im vorgestellten Modell jedoch nicht identisch: Die Transkription kann durch unterschiedliche Regulatoren beeinflusst werden und durch verschiedene Produktionsfunktionen erfolgen.

Beispiel 1 (Genom und induzierter GRN-Graph) *Dieses Beispiel soll die eben beschriebene Abbildung eines Genoms auf einen GRN-Graph verdeutlichen. Außerdem sollen der Unterschied zwischen multiplen und einfachen GRN-Graph sowie die eben besprochene gemeinsame Nutzung von Regulatoren illustriert werden.*

Gegeben sei eine Menge von RNS $\mathrm{S_{RNS}} = \{m, n, o\}$*, eine Menge von Proteinen* $\mathrm{S_{Prot}} = \{Q_1, Q_2\}$ *und es gebe ein Dimer* $\langle Q_1, Q_2\rangle$*. Damit gilt:* $\mathrm{S} = \{m, n, o, Q, R, \langle Q_1, Q_2\rangle\}$*. Weiterhin sei ein Genom* $\mathfrak{G} = \{A, B, C, D\}$ *mit folgenden Zuordnungen gegeben: A und B besitzen das Primärtranskript m, C besitzt n und D besitzt o und*

$$P(m) = Q_1, \tag{4.12}$$
$$P(n) = Q_2, \tag{4.13}$$
$$P(o) = \bot\,. \tag{4.14}$$

Die entsprechenden Produktionseinheiten seien:

$$\Pi_A = (m, \emptyset, \mathrm{Prd}_A) \tag{4.15}$$
$$\Pi_B = (m, \{\langle Q_1, Q_2\rangle\}, \mathrm{Prd}_B) \tag{4.16}$$
$$\Pi_C = (n, \emptyset, \mathrm{Prd}_C) \tag{4.17}$$
$$\Pi_D = (o, \{Q_1\}, \mathrm{Prd}_D) \tag{4.18}$$
$$\Pi_m = (Q_1, \{Q_2\}, \mathrm{Prd}_m) \tag{4.19}$$
$$\Pi_n = (Q_2, \emptyset, \mathrm{Prd}_n) \tag{4.20}$$

mit geeigneten Produktionsfunktionen $\mathrm{Prd}_i, i \in \{A, B, C, D, m, n\}$.

In Abbildung 4.5 sind das Beispiel und die zugehörigen GRN-Graphen dargestellt. Die gestrichelte Kanten sind nur im multiplen GRN-Graphen vorhanden.

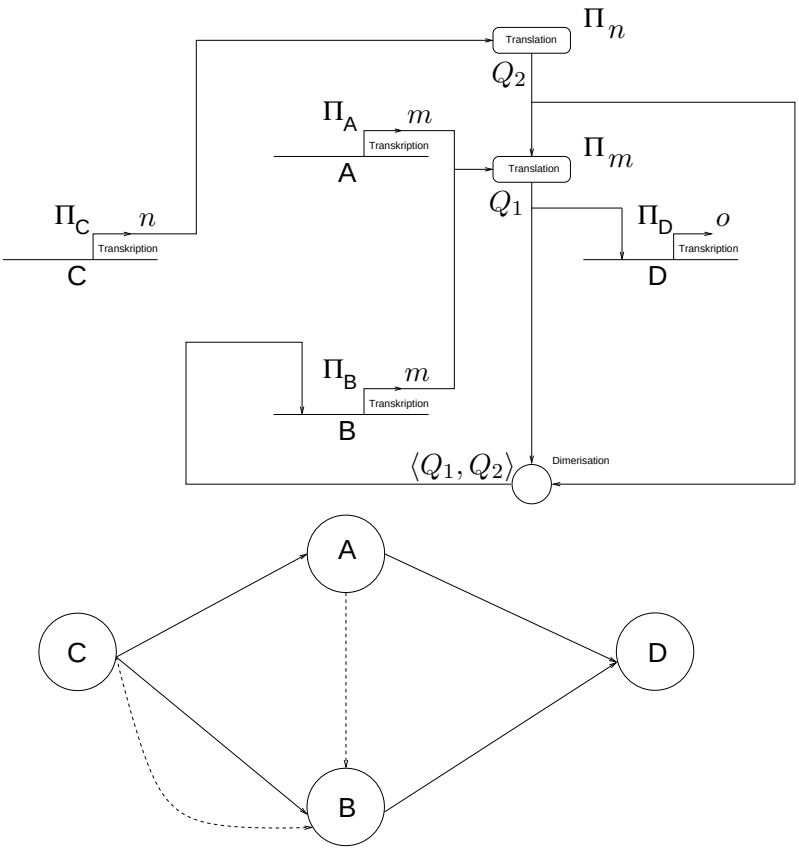

Abb. 4.5.: Genom $\mathfrak{G} = \{A, B, C, D\}$ aus Beispiel 1 (oben) und zugehöriger einfacher bzw. multipler GRN-Graph (unten). Durchgezogene Linie sind gemeinsame Kanten im multiplen und einfachen GRN-Graph, gestrichelte Linien sind nur im multiplen GRN-Graphen vorhanden.

4.1.4. Eigenschaften

Nicht alle Regulatoren, die in einem Genom einen Einfluss ausüben, müssen von Genen produziert werden:

Definition 4.13 (externer Regulator)

Ein Regulator $s \in \mathrm{R}(G)$ für ein Gen $G \in \mathfrak{G}$ heißt **extern** *genau dann, wenn gilt:*

$$\mathfrak{G}_{\mathrm{Prod}}(s) = \emptyset. \tag{4.21}$$

Falls ein Gen eine seiner Regulatorsubstanzen produziert, übt es einen direkten Einfluss auf sich selbst aus:

Definition 4.14 (selbstregulierend)

Ein Gen $G \in \mathfrak{G}$ heißt **selbstregulierend**, *falls es einen Regulator $s \in \mathrm{R}(G)$ gibt, der von G produziert wird, für den also gilt: $G \in \mathfrak{G}_{\mathrm{Prod}}(s)$.*

Im Allgemeinen gibt der Eingangsgrad eines Objektes an, wieviele andere Objekte einen direkten Einfluss auf es ausüben. In der vorliegenden Modellierung sind unterschiedliche Auffassungen vom Eingangsgrad (eines Gens) möglich: Der einfachste Fall zählt die Anzahl von Genen, die einen direkten Einfluss ausüben. Sollen die verschiedenen Wege berücksichtigt werden, auf denen ein Gen ein anderes beeinflusst, erhält man einen anderen, im Allgemeinen größeren Eingangsgrad. Dabei wird jeder direkte regulierende Einfluss gezählt, den Gene auf ein Gen ausüben. Schließlich ist eine Auffassung aus Sicht der Produktionseinheiten möglich: Wieviele Substanzen üben einen direkten Einfluss auf die Produktion aus? Dies entspricht der Anzahl der Regulatoren.

Definition 4.15 (Eingangsgrad)

Für ein Gen $G \in \mathfrak{G}$ mit zugehörigem Primärtranskript $m \in \mathrm{S}_{\mathrm{RNS}}$ und Transkriptionseinheit $\Pi_G = (m, \mathrm{R}_G, \mathrm{Prd}_G)$ werden folgende Eingangsgrade definiert:

(15–1) *Der* **regulatorische Eingangsgrad** *k^r_G von G ist die Anzahl der transkriptionalen und posttranskriptionalen Regulatoren von G:*

$$k^r_G \;=_{df}\; \begin{cases} \mathrm{card}\big(\mathrm{R}_G\big) + \mathrm{card}\big(\mathrm{R}_m\big) & \textit{falls } P(m) \neq \perp, \\ \mathrm{card}\big(\mathrm{R}_G\big) & \textit{sonst}, \end{cases} \tag{4.22}$$

wobei für kodierendes m die Translationseinheit $\Pi_m = (P(m), \mathrm{R}_m, \mathrm{Prd}_m)$ sei.

(15–2) *Der* **einfache Eingangsgrad** *k_G von G ist die Anzahl von Genen, die Regulatoren von G produzieren:*

$$k_G \;=_{df}\; \mathrm{card}\Big(\bigcup_{s \in \mathrm{R}(G)} \mathfrak{G}_{\mathrm{Prod}}(s)\Big). \tag{4.23}$$

(15–3) *Der* **multiple Eingangsgrad** *k^m_G von G ist die Summe der Anzahl von Produzenten aller transkriptionalen und posttranskriptionalen Regulatorsubstanzen von G:*

$$k^m_G \;=_{df}\; \sum_{s \in \mathrm{R}_G} \mathit{card}\big(\mathfrak{G}_{\mathrm{Prod}}(s)\big) + \sum_{s \in \mathrm{R}_m} \mathit{card}\big(\mathfrak{G}_{\mathrm{Prod}}(s)\big), \tag{4.24}$$

wobei für kodierendes m die Translationseinheit $\Pi_m = (P(m), \mathrm{R}_m, \mathrm{Prd}_m)$ sei und für nicht kodierendes m gelte $\mathrm{R}_m =_{df} \emptyset$.

Das die Definition des einfachen bzw. des multiplen Eingangsgrads im Zusammenhang mit den Eingangsgraden im induzierten einfachen bzw. multiplen GRN-Graph steht, zeigt folgender Satz:

Satz 4.1

(1) Für ein Gen $G \in \mathfrak{G}$ entspricht der einfache Eingangsgrad der Anzahl eingehender Kanten von G im von $\mathfrak{G}$ induzierten einfachen GRN-Graphen.

(2) Für ein Gen $G \in \mathfrak{G}$ entspricht der multiple Eingangsgrad der Anzahl eingehender Kanten von G im von $\mathfrak{G}$ induzierten multiplen GRN-Graphen.

Der Satz folgt direkt aus den Definitionen von GRN-Graph und Eingangsgrad.

Der folgende Satz fasst einige leicht zu verifizierende Beziehungen zwischen den verschiedenen Auffassungen von Eingangsgrad zusammen:

Satz 4.2

(1) Für jedes $G \in \mathfrak{G}$ gilt:

$$k_G \quad \leq \quad k_G^m. \tag{4.25}$$

(2) Für ein Gen $G \in \mathfrak{G}$ mit zugehörigem Primärtranskript $m \in \mathrm{S}_{\mathrm{RNS}}$ und Transkriptionseinheit $\Pi_G = (m, \mathrm{R}_G, \mathrm{Prd}_G)$ gilt

$$k_G = k_G^m \tag{4.26}$$

genau dann, wenn die Produzentenmengen aller direkten Regulatoren von G paarweise disjunkt sind und kein direkter Regulator sowohl transkriptionaler als auch posttranskriptionaler Regulator von G ist:

$$\forall s, q \in \mathrm{R}(G) : s \neq q \rightarrow \mathfrak{G}_{\mathrm{Prod}}(s) \cap \mathfrak{G}_{\mathrm{Prod}}(q) = \emptyset \quad \wedge \quad \mathrm{R}_G \cap \mathrm{R}_m = \emptyset, \tag{4.27}$$

wobei für kodierendes m die Translationseinheit $\Pi_m = (P(m), \mathrm{R}_m, \mathrm{Prd}_m)$ sei und für nicht kodierendes m gelte $\mathrm{R}_m =_{df} \emptyset$.

(3) Für ein Gen $G \in \mathfrak{G}$, das keine externen Regulatoren besitzt, gilt:

$$k_G^r \leq k_G^m. \tag{4.28}$$

(4) Für ein Gen $G \in \mathfrak{G}$, das keine externen Regulatoren besitzt, ist

$$k_G^r = k_G^m \tag{4.29}$$

genau dann, wenn für alle $s \in \mathrm{R}(G)$ gilt: $\mathrm{card}(\mathfrak{G}_{\mathrm{Prod}}(s)) = 1$.

4.1.5. Dynamik

Das bis hierhin entwickelte Modell beschreibt den Zustand eines Systems und legt die Prozesse fest, die diesen Zustand ändern können. Außerdem wurde gezeigt, wie ein solches Modell auf die klassische Darstellung eines Genregulationsnetzes abgebildet werden kann. Bisher wurde noch nichts darüber gesagt, wie die Prozesse den Zustand des Systems ändern. Die im folgenden definierte Dynamik vervollständigt die bisher strukturelle Beschreibung des Modells.

Wie oben dargelegt, kann sich der Zustand des Systems durch folgenden Prozesse ändern:

1. Bei der Bildung einer zusammengesetzten Substanz nimmt deren Konzentration zu.
2. Beim Verbrauch von Substanzen durch Bildung einer zusammengesetzten Substanz kommt es zu einer Konzentrationsabnahme der bildenden Substanzen.
3. Der Zerfall bewirkt eine beständige Konzentrationsabnahme aller Substanzen.
4. Die externe Änderung kann sowohl eine Konzentrationszu- als auch eine -abnahme hervorrufen.
5. Die Produktionseinheiten produzieren einfache Substanzen, führen also zu einer Konzentrationszunahme dieser Substanzen.

Die folgende Definition beschreibt die Anteile der einzelnen Prozesse an der Zustandsänderung des Systems.

Definition 4.16 (Konzentrationsänderung)

(16–1) bei Bildung *(Combination)*
Für jede zusammengesetzte Substanz $s = \langle s, q\rangle \in \mathrm{S}_{i+1}$ *mit minimalen* i *wird die Bildungsrate* $(dx_{\langle s,q\rangle}/dt)_{\mathrm{comb}}$ *definiert:*

$$\left(\frac{dx_{\langle s,q\rangle}}{dt}\right)_{\mathrm{comb}} =_{df} \quad \gamma_i(s,q) \times x_s(t) \times x_q(t) \geq 0. \tag{4.30}$$

(16–2) durch Verbrauch bei Bildung
Für jede zusammengesetzte Substanz $\langle s, q\rangle \in \mathrm{S}_{i+1}$ *mit minimalem* i *werden die Verbrauchsraten* $(dx_s/dt)_{\langle s,q\rangle}$ *und* $(dx_q/dt)_{\langle s,q\rangle}$ *definiert. Für* $s \neq q$*:*

$$\left(\frac{dx_s}{dt}\right)_{\langle s,q\rangle} = \left(\frac{dx_q}{dt}\right)_{\langle s,q\rangle} =_{df} \quad -\gamma_i(s,q) \times x_s(t) \times x_q(t) \leq 0, \tag{4.31}$$

und für $s = q$*:*

$$\left(\frac{dx_s}{dt}\right)_{\langle s,s\rangle} = \quad =_{df} \quad -2\gamma_i(s,s) \times x_s(t)^2 \leq 0. \tag{4.32}$$

(Der Faktor 2 kommt aus der Stöchiometrie der Reaktionsgleichung $2s \rightarrow s_2$*).*

(16–3) durch Zerfall *(Decay)*
Für jede Substanz $s \in \mathrm{S}$ *ist die zeitliche Änderung der Konzentration durch Zerfall* $(dx_s/dt)_{\mathrm{decay}}$ *gegeben als:*

$$\left(\frac{dx_s}{dt}\right)_{\mathrm{decay}} =_{df} \quad -\delta_s x_s(t) \leq 0. \tag{4.33}$$

(16–4) durch externe Änderung
Für jede Substanz $s \in \mathrm{S}$ *ist für jeden Zeitpunkt* t *die externe Änderungsrate gegeben durch:*

$$\left(\frac{dx_s}{dt}\right)_{\mathrm{ext}} =_{df} y_s(t). \tag{4.34}$$

(16–5) durch Produktion

Für jede Transkriptionseinheit $\Pi = (m, \mathrm{R}, \mathrm{Prd})$ *ergibt sich die Änderungsrate* $(dx_m/dt)_\Pi$ *der Konzentration von* $m \in \mathrm{S_{RNS}}$ *bezüglich* Π *in Abhängigkeit der Konzentrationen der Regulatoren:*

$$\left(\frac{dx_m}{dt}\right)_\Pi =_{df} \mathrm{Prd}(x_{r_1}(t), \ldots, x_{r_k}(t)) \geq 0, \tag{4.35}$$

wobei $\mathrm{R} = \{r_1, \ldots, r_k\}$ *gelte.*

Für jede Translationseinheit $\Pi = (P(m), \mathrm{R}, \mathrm{Prd})$ *ergibt sich die Änderungsrate* $\left(dx_{P(m)}/dt\right)_\Pi$ *der Konzentration von* $P(m) \in \mathrm{S_{Prot}}$ *bezüglich* Π *in Abhängigkeit der Konzentrationen der Regulatoren und von* $m \in \mathrm{S_{RNS}}$*:*

$$\left(\frac{dx_{P(m)}}{dt}\right)_\Pi =_{df} \mathrm{Prd}(x_{r_1}(t), \ldots, x_{r_k}(t), x_m) \geq 0, \tag{4.36}$$

wobei $\mathrm{R} = \{r_1, \ldots, r_k\}$ *gelte.*

Die folgende Definition gibt die Bedingungen an, unter denen ein Genom (und die zugehörige Menge von Produktionseinheiten) ein Genregulationssystem bilden. Die Änderungsrate der Konzentration jeder Substanz ist die Summe der Änderungsraten aller auf sie einwirkenden Prozesse. In Abbildung 4.6 sind die Substanzströme für eine einfache und eine zusammengesetzte Substanz beispielhaft veranschaulicht.

Definition 4.17 (Genregulationssystem)

Gegeben sei ein Genom $\mathfrak{G} = \{G_1, \ldots, G_N\}$*, bestehend aus* N *Genen.* $\mathfrak{G}$ *heißt* **Genregulationssystem** *genau dann, wenn folgende Eigenschaften gelten:*

(17–1) *Für alle einfachen Substanzen* $s \in \mathrm{S_0}$ *ist die zeitliche Änderungsrate der Konzentration* dx_s/dt *gegeben durch:*

$$\frac{dx_s}{dt} = \sum_{\Pi=(s,\mathrm{R},\mathrm{Prd})} \left(\frac{dx_s}{dt}\right)_\Pi + \sum_{\langle s,q\rangle \in \mathrm{S_1}} \left(\frac{dx_s}{dt}\right)_{\langle s,q\rangle} + \left(\frac{dx_s}{dt}\right)_{\mathrm{ext}} + \left(\frac{dx_s}{dt}\right)_{\mathrm{decay}}. \tag{4.37}$$

(17–2) *Für alle zusammengesetzten Substanzen* $s \in \mathrm{S} \setminus \mathrm{S_0}$ *ist die zeitliche Änderungsrate der Konzentration* dx_s/dt *gegeben durch:*

$$\frac{dx_s}{dt} = \left(\frac{dx_s}{dt}\right)_{\mathrm{comb}} + \sum_{\langle s,q\rangle \in \mathrm{S}} \left(\frac{dx_s}{dt}\right)_{\langle s,q\rangle} + \left(\frac{dx_s}{dt}\right)_{\mathrm{ext}} + \left(\frac{dx_s}{dt}\right)_{\mathrm{decay}}. \tag{4.38}$$

Für eine konkrete Ausprägung eines Genregulationssystems müssen die Bildungsraten für jede zusammengesetzte Substanz, die Zerfallsrate und die externe Änderung aller Substanzen sowie die Produktionsfunktionen festgelegt werden. Die Verbindungsstruktur wird über die Regulatoren der Produktionseinheiten bestimmt. In den folgenden Abschnitten wird gezeigt, wie eine solche konkrete Ausprägung erfolgen kann.

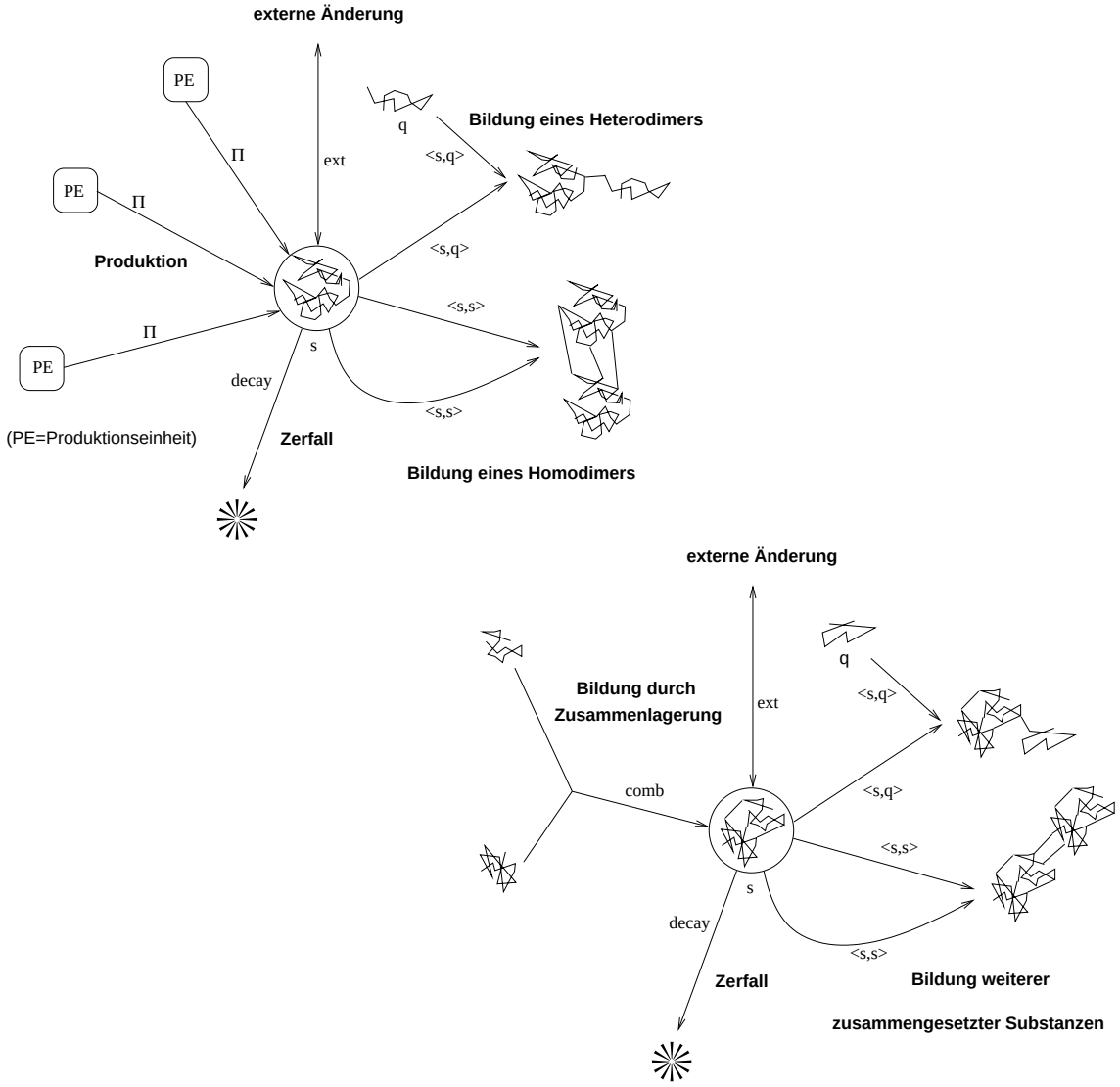

Abb. 4.6.: Substanzströme für eine einfache (oben) und eine zusammengesetzte (unten) Substanz. Die Kanten sind mit den Indizes der Änderungsraten in Definition 4.16 beschriftet.

4.2. Erzeugung künstlicher Genregulationssysteme

In diesem Abschnitt wird beschrieben, wie künstliche Genregulationssysteme entsprechend der vorgestellten Modellierung erzeugt werden können. Das Ziel besteht in der zufälligen Erzeugung von Systemen, die bestimmte Eigenschaften besitzen. Durch entsprechende Parameter können die gewünschten Eigenschaften an die Bedürfnisse angepasst werden. Die Erzeugung erfolgt in zwei Stufen: erst wird die Verbindungsstruktur durch Wahl der Produkte und Regulatoren der Gene aufgebaut (4.2.1) und anschließend ein zugehöriges Differenzialgleichungssystem abgeleitet (4.2.2). Die Umsetzung des hier beschriebenen Verfahrens in ein Computerprogramm wird im letzten Abschnitt (4.3) skizziert.

Für die Erzeugung einer zufälligen Verbindungsstruktur eines künstlichen Netzes der Größe N sind folgende Ansätze denkbar:

1. Knoten werden entsprechend einer bestimmten Verteilung zufällig miteinander verbunden. Die Verteilung kann z.B. den Eingangs- oder den Ausgangsgrad betreffen. Diese Methode wird bei den in Abschnitt 3.2.1 beschriebenen, zufällig erzeugten Genregulationsnetzen und auch bei dem hier implementierten Verfahren (siehe unten) verwendet.

2. Die Modellierung der Genexpression durch künstliche Genome, wie sie von z.B. Reil (1999) und Eggenberger (1997) benutzt werden, lässt sich ebenfalls für die Erzeugung künstlicher Verbindungsstrukturen verwenden. Anstatt Knoten willkürlich miteinander zu verbinden, werden die Kanten durch Mustervergleiche auf dem zugrundeliegenden künstlichen Genom bestimmt. Dieses Vorgehen ist näher an den biologischen Prozessen, wo Regulation auf Bindung von Substanzen an geeigneten Stellen beruht. Eventuell lassen sich somit interessantere, d.h. biologisch plausiblere, Netzstrukturen entwickeln.

3. (Kauffman, 1993, S. 419ff) beschreibt eine Methode, mit der sich Netze „züchten“ lassen. Er geht ebenfalls von einem künstlichen Genom aus, jedoch mit einer zugehörigen, regelmäßigen Verbindungsstruktur (z.B. isolierte Knoten mit Selbstregulation bzw. ein großer Regulationskreislauf). Auf diesem künstlichen Genom werden zufällige Mutationen (Vertauschen und Vervielfältigen von Abschnitten) vorgenommen. Vorstellbar wäre dieses Vorgehen kombiniert mit einem genetischen Algorithmus. Einen geeigneten Selektionsdruck vorausgesetzt, könnten so eventuell Netze mit einer ähnlichen Struktur wie biologische Genregulationsnetze erhalten werden.

4.2.1. Verfahren zur Erzeugung eines künstlichen Genregulationsnetzes

Das im Folgenden beschriebene, eigene Verfahren erzeugt ein Genom mit N Genen und eine zugehörige Menge von Produktionseinheiten. Durch die Parameter in Tabelle 4.2 können die entsprechenden Eigenschaften des erzeugten Systems vorgegeben werden. Einige der Eigenschaften können vom Algorithmus nur näherungsweise erfüllt werden, diese sind in der Tabelle mit * gekennzeichnet. So kann zum Beispiel der Anteil von Selbstregulatoren durch das Verfahren nicht garantiert werden: Die oben besprochene gemeinsame Nutzung von Regulatoren zwischen Genen schränkt die Möglichkeiten bei der Auswahl von Regulatorsubstanzen zur Etablierung eines selbstregulierenden Gens ein. Die vom Verfahren erzeugten Verbindungsstrukturen sollen in bestimmten Elementen zufällig sein. Aus diesem Grund werden einige Auswahlen im Algorithmus zufällig getroffen. Dies verhindert eventuell eine für die Garantierung einer Eigenschaft optimale Wahl.

Im abstrakten Modell (Kapitel 4.1) wird die Wirkung *aller* Substanzen auf die Regulation der Transkription bzw. Translation zugelassen. In welchem Maße jedoch z.B. RNS an der Regulation beteiligt ist, ist eine offene Frage. Im klassischen Modell der transkriptionalen Regulation spielen Transkriptionsfaktoren und andere Proteine die entscheidende Rolle (vergleiche z.B. Munk (2001)). Bei der posttranskriptionalen Regulation gibt es immerhin die Möglichkeit der Blockierung der mRNS durch komplementäre RNS („Antisense-RNS“, (Munk, 2001, S. 10-19)). Neuere Untersuchungen zeigen, dass bestimmte RNS-Sorten (doppelsträngige RNS) auch auf die Transkription einen Einfluss ausüben („RNA interference (RNAi)“, Hannon (2002)). Um dieser strittigen Frage gerecht zu werden, wird

Tab. 4.2.: Parameter für das Verfahren. Mit * gekennzeichnete Eigenschaften werden vom Verfahren nur in einem statistischen Sinne garantiert.

Parameter	**Symbol**	**Beschreibung**	**Wertebereich**
Größe	N	Anzahl der Gene	$\mathbb{N}_{>0}$
Anzahl RNS	n_{rns}		$\mathbb{N}_{>0}$
Anzahl Proteine	n_{prot}		$\mathbb{N}$
Anteil nicht kodierender RNS	$\alpha_{\perp}$	Anteil der RNS, die nicht kodierend sein soll	$[0,1]$
Anzahl Homodimere	n_{hom}		$\mathbb{N}$
Anzahl Heterodimere	n_{het}		$\mathbb{N}$
Dimerisationsneigung von RNS	γ_{rns}	Neigung von RNS; Dimere zu bilden	$[0,1]$
Anzahl *Hubs**	n_{hub}	Anzahl der Gene mit hohem Ausgangsgrad	$\mathbb{N}$
Anteil Selbstregulatoren*	α_{self}	Anteil der Gene, die sich selbst regulieren	$[0,1]$
Anteil externer Regulatoren*	α_{ext}	Anteil der Regulatorsubtanzen, die nicht Produkt eines Genes sind	$[0,1]$
maximaler multipler Eingangsgrad	$k^m_{\max}$		$\{1,\ldots,N-1\}$
Eingangsgradverteilung*	$K^m = \{\omega_0,\ldots,\omega_{k^m_{\max}}\}$	diskrete Wahrscheinlichkeitsverteilung, die zu jedem multiplen Eingangsgrad i $(0 \leq i \leq k^m_{\max})$ eine Wahrscheinlichkeit $\omega(k^m = i) = \omega_i$ angibt	$[0,1]^{k^m_{\max}+1}$
Regulationsneigung von RNS*	ρ_{rns}	Neigung von RNS, regulierend zu wirken	$[0,1]$
Anteil der transkriptionalen Regulation*	α_{tkr}	transkriptionaler Anteil des multiplen Eingangsgrads	$[0,1]$

durch einen entsprechenden Parameter (ρ_{rns}) ermöglicht, die Neigung von RNS (und RNS-Dimeren), sich an der Regulation zu beteiligen, festzulegen. Setzt man diesen Parameter auf den Wert 0.0, verhindert man jeglichen regulativen Einfluss durch RNS. Durch einen Wert von 1.0 besitzen RNS und Proteine dieselbe Wahrscheinlichkeit, einen regulativen Einfluss auszuüben.

Das abstrakte Modell ermöglicht die Bildung jeglicher Art von zusammengesetzten Substanzen. In welchem Maße RNS an der Bildung von Dimeren oder komplexeren Gebilden in der Zelle beteiligt ist, ist eine offene Frage. Zwar gibt es doppelsträngige RNS (Hannon (2002)), ob jedoch auch Dimere zwischen Proteinen und RNS eine Rolle spielen, ist unklar. Durch einen entsprechenden Parameter (γ_{rns}) kann die Neigung von RNS, zusammengesetzte Substanzen zu bilden, bestimmt werden.

Die zentralen Parameter des Verfahrens sind die Anzahl N der Gene (Größe des Genoms) und die multiple Eingangsgradverteilung K^m. Durch zwei Parameter kann ein unterschiedliches Verhalten von RNS und Proteinen bei der Dimerisation (γ_{rns}) und der Regulation (ρ_{rns}) bewirkt werden (siehe oben). Durch α_{tkr} kann bestimmt werden, welcher Vorgang (Transkription oder Translation) regulatorisch stärker beeinflusst wird.

Die Gesamtanzahl an Substanzen in einem künstlichen Genregulationssystem sollte etwa das zwei- bis fünffache der Anzahl der Gene betragen. Eine geringere Anzahl von Substanzen bewirkt eine hohe Anzahl von Produzenten für jede Substanz und erschwert so die Realisierung kleiner multipler Eingangsgrade. Eine Substanzanzahl, die um ein Vielfaches höher als die Anzahl der Gene ist, führt zu einer kleinen Wahrscheinlichkeit, dass eine Substanz mehrere Produzenten besitzt. Ein hoher multipler Eingangsgrad kann dann nur durch einen hohen regulatorischen Eingangsgrad realisiert werden.

Realisierung

Die Grundidee des Verfahrens besteht in einer Umsetzung von Definition 4.12. Nachdem die Produkte für jedes Gen festgelegt wurden, kann ein sequentieller Aufbau einer Verbindungsstruktur erfolgen: Für jedes Gen werden geeignete Regulatoren gewählt, die einen bestimmten Eingangsgrad und Selbstregulation bzw. nicht Selbstregulation erfüllen. Zuerst müssen dafür die Mengen der Substanzen und Gene festgelegt werden. Das Verfahren besteht aus folgenden Schritten:

1. Festlegung der Menge der RNS und Proteine (Algorithmus 4.1).
2. Bildung zusammengesetzter Substanzen aus RNS und Proteinen (Algorithmus 4.2). Es werden nur zusammengesetzte Substanzen der 1. Stufe (Dimere) betrachtet , es gilt also $\mathrm{S} = \mathrm{S}_1$.
3. Konstruktion der Symbolmenge der Gene: $\mathfrak{G} := \{G_1, \ldots, G_N\}$ und Wahl des multiplen Eingangsgrads für jedes Gen und Aufteilung auf transkriptionale und posttranskriptionale Regulation. Dabei wird für jedes Gen ein geeignetes Primärtranskript $m \in \mathrm{S}_{\mathrm{RNS}}$ gewählt (Algorithmus 4.4).
4. Bestimmung der Produzenten für jede Substanz $s \in \mathrm{S}$ (Algorithmus 4.5).
5. Auswahl von n_{hub} Regulatorsubstanzen $H := \{r_1, \ldots, r_{n_{\mathrm{hub}}}\}$, die jeweils genau einen Produzenten besitzen und deren Produzenten paarweise verschieden sind. Diese Produzenten repräsentieren *Hubs*, das heißt Gene, die einen sehr hohen Ausgangsgrad besitzen. Dies wird erreicht, indem die Substanzen aus H bevorzugt als Regulatorsubstanzen im Schritt 6 ausgewählt werden.
6. Konstruktion der Regulatormengen für jede Produktionseinheit (Algorithmus 4.6).

Wenn in den folgenden Algorithmen eine zufällige Wahl aus einer Menge ohne eine weitere Erklärung erfolgt, handelt es sich um eine Auswahl nach einer Gleichverteilung: Jedes Objekt der Menge wird mit gleicher Wahrscheinlichkeit gezogen. Die Parameter in Tabelle 4.2 sowie die im Folgenden definierten Mengen seien den Algorithmen global bekannt, auch wenn sie nicht explizit als Eingabe gekennzeichnet sind.

Die Mengen der einfachen Substanzen werden durch Wahl einer geeigneten Anzahl von Symbolen festgelegt. Danach wird für jede RNS ein Protein bzw. $\perp$ (entsprechend $\alpha_\perp$) als Produkt gewählt (Algorithmus 4.1).

Algorithmus 4.1 (Festlegung der einfachen Substanzen und der Produktfunktion)

Eingabe: n_{rns} *Anzahl RNS*
n_{prot} *Anzahl Proteine*
$\alpha_\perp \in [0,1]$ *Anteil nicht kodierender RNS*
Ausgabe: S_{Prot} *Symbolmenge der Proteine*
S_{RNS} *Symbolmenge der RNS*
$P : \mathrm{S}_{\text{RNS}} \to \mathrm{S}_{\text{Prot}} \cup \{\perp\}$ *Produktfunktion*

$\mathrm{S}_{\text{Prot}} := \{p_1, \ldots, p_{n_{\text{prot}}}\}$; *Wahl der Protein-*
$\mathrm{S}_{\text{RNS}} := \{m_1, \ldots, m_{n_{\text{rns}}}\}$; *und der RNS-Symbole*

for all $m \in \mathrm{S}_{\text{RNS}}$ **do** *Festlegung der Produktfunktion*
Setze $P(m) := p_i \in \mathrm{S}_{\text{Prot}}$ mit Wahrscheinlichkeit

$$\omega(P(m) = p_i) = \frac{1 - \alpha_\perp}{n_{\text{prot}}}$$

bzw. $P(m) := \perp$ mit

$$\omega(P(m) = \perp) = \alpha_\perp;$$

end for

Aus diesen einfachen Substanzen können die zusammengesetzten Substanzen aufgebaut werden. Das Verfahren beschränkt sich dabei auf die Bildung von Dimeren. Eine Erweiterung auf komplexere Strukturen ist jedoch möglich. Die Homo- und Heterodimere werden aus zufällig gewählten Proteinen bzw. RNS zusammengesetzt. Durch den Parameter α_{rns} kann festgelegt werden, ob Dimere genauso stark zur Dimerisation neigen wie Proteine ($\alpha_{\text{rns}} = 1$) oder weniger stark ($\alpha_{\text{rns}} < 1$) bzw. gar keine Dimere bilden ($\alpha_{\text{rns}} = 0$). Die Neigung γ_0 zur Bildung eines Dimers wird gleichverteilt aus einem vorgegebenen Intervall I_γ gewählt. Diese Intervall umfasst nach Definition 4.3 die positiven reellen Zahlen. In einer konkreten Implementierung des Verfahrens wird ein Teilintervall benutzt (siehe Tabelle 4.3).

Algorithmus 4.2 (Definition der Menge der Substanzen)

Eingabe: n_{het} *Anzahl Heterodimere*
n_{hom} *Anzahl Homodimere*
$\alpha_{\text{rns}} \in [0,1]$ *Beteiligung von RNS an Dimeren*
Ausgabe: S *Menge der Substanzen*

for $i := 1$ to n_{hom} **do** *Homodimere*
$s :=$ chooseSimpleSubstance(); *Algorithmus 4.3*
Wahl von $\gamma_0\langle s, s\rangle$ gleichverteilt aus I_γ; *Neigung zur Bildung von* $\langle s, s\rangle$
end for
for $i := 1$ to n_{het} **do** *Heterodimere*
repeat
$s :=$ chooseSimpleSubstance(); *Algorithmus 4.3*
$q :=$ chooseSimpleSubstance();
until $s \neq q$ *Heterodimer!*
Wahl von $\gamma_0\langle s, q\rangle$ gleichverteilt aus I_γ; *Neigung zur Bildung von* $\langle s, q\rangle$
end for
$\mathrm{S}_1 := \mathrm{S}_0 \cup \{\langle s, q\rangle : s, q \in \mathrm{S}_0 \wedge \gamma_0\langle s, q\rangle > 0\}$; *Dimere und einfache Substanzen*
$\gamma_1(s, q) := 0$ für alle $s, q \in \mathrm{S}_1$ *damit gilt:* $\mathrm{S} = \mathrm{S}_1$

Die Wahl einer Substanz erfolgt mit einer Hilfsfunktion:

Algorithmus 4.3 (Hilfsfunktion zur Auswahl einer einfachen Substanz)

function s = chooseSimpleSubstance()
  $\omega(\mathsf{RNS}) := \alpha_{\mathrm{rns}} \times \frac{n_{\mathrm{rns}}}{n_{\mathrm{rns}}+n_{\mathrm{prot}}}$; *Wahrscheinlichkeit, eine RNS zur Dimerisation auszuwählen*
  Wähle zufällig eine Substanz s mit Wahrscheinlichkeit $\omega(\mathsf{RNS})$ aus $\mathrm{S}_{\mathrm{RNS}}$ oder mit $(1-\omega(\mathsf{RNS}))$ aus $\mathrm{S}_{\mathrm{Prot}}$;
end function

Als nächster Schritt erfolgt die Festlegung der Symbolmenge $\mathfrak{G} = \{G_1, \ldots, G_N\}$ der Gene und die Wahl eines multiplen Eingangsgrads k_G^m für jedes Gen $G \in \mathfrak{G}$ entsprechend der multiplen Eingangsgradverteilung K^m. Dieser wird nach der Vorgabe α_{tkr} auf transkriptionale und posttranskriptionale (translationale) Regulation aufgeteilt. Bei der Aufteilung werden jedoch Abweichungen zugelassen: Der Anteil transkriptionaler Regulation k_G^{tkr} wird normalverteilt um den Wert $\alpha_{\mathrm{tkr}} \times k_G^m$ gewählt. Die Standardabweichung σ ist dabei am größten ($\sigma = k_G^m/4$), wenn transkriptionale und translationale Regulation ausgeglichen sein sollen ($\alpha_{\mathrm{tkr}} = 0.5$) und 0, falls eine von beiden Regulationsarten ausschließlich erfolgen soll ($\alpha_{\mathrm{tkr}} \in \{0, 1\}$). Dabei ist zu beachten, dass Gene, die dasselbe Primärtranskript besitzen, auch dieselbe Translationseinheit und damit denselben translationalen Anteil des multiplen Eingangsgrads aufweisen. Aus diesem Grund wird an dieser Stelle erst ein Primärtranskript $m \in \mathrm{S}_{\mathrm{RNS}}$ für jedes Gen gewählt. Dies kann entweder eine RNS sein, die noch von keinem anderen Gen produziert wird oder eine RNS, deren produzierende Gene den gewünschten translationalen Anteil am multiplen Eingangsgrad besitzen. Das Verfahren stellt sicher, dass alle Gene, die dasselbe Primärtranskript besitzen, denselben translationalen Anteil am multiplen Eingangsgrad besitzen.

Algorithmus 4.4 (Erzeugung des Genoms mit zugeordneten multiplen Eingangsgraden und Primärtranskripten)

Eingabe: K^m *Sollverteilung des multiplen Eingangrads*
  α_{tkr} *Anteil transkriptionaler Regulation*
Ausgabe: $\mathfrak{G}$ *Symbolmenge der Gene (Genom)*
  $(k_G^m)_{G\in\mathfrak{G}}$ *(gesamte) multiple Eingangsgrade*
  $(k_G^{\mathrm{tkr}})_{G\in\mathfrak{G}}$ *transkriptionale Anteile der multiplen Eingangsgrade*
  $(m)_{G\in\mathfrak{G}}$ *Primärtranskripte*

$\mathfrak{G} = \{G_1, \ldots, G_N\}$ *Wahl der Gen-Symbole*
for all $G \in \mathfrak{G}$ **do**
  repeat
    Wähle k_G^m entsprechend K^m;
    $\overline{x} := \frac{k_G^m}{2}$; *Mittelwert*
    $\sigma := k_G^m \times \alpha_{\mathrm{tkr}} \times (1-\alpha_{\mathrm{tkr}})$; *Standardabweichung*
    repeat
      Wähle k_G^{tkr} normalverteilt um $\overline{x}$ mit Standardabweichung σ;
    until $0 \le k_G^{\mathrm{tkr}} \le k_G^m$ *zulässiger Wert für k_G^{tkr}?*
    Wähle $m \in \mathrm{S}_{\mathrm{RNS}}$;
    if ist m schon Primärtranskript eines Gens G'? **then**
      $k_G^{\mathrm{trl}} := k_{G'}^{\mathrm{trl}}$; *gleicher translationaler Anteil*
    else
      $k_G^{\mathrm{trl}} := k_G^m - k_G^{\mathrm{tkr}}$;
    end if
  until $k_G^{\mathrm{trl}} = k_G^m - k_G^{\mathrm{tkr}}$ *geeignetes Primärtranskript gefunden?*
end for

Da für jedes Gen ein Primärtranskript und damit implizit über die Produktfunktion ein eventuell produziertes Protein festgelegt wurden, kann jetzt für jede Substanz $s \in \mathrm{S}$ die Menge der Produzenten entsprechend Definition 4.11 bestimmt werden:

Algorithmus 4.5 (Bestimmung der Produzenten einer Substanz)
Eingabe: $s \in \mathrm{S}$
Ausgabe: $\mathfrak{G}_{\mathrm{Prod}}(s) \subseteq \mathfrak{G}$ *Produzenten von s*

if $s \in \mathrm{S}_{\mathrm{RNS}}$ **then**
  $\mathfrak{G}_{\mathrm{Prod}}(s) := \{G \in \mathfrak{G} : \text{s ist Primärtranskript von } G\}$;
else if $s \in \mathrm{S}_{\mathrm{Prot}}$ **then**
  $\mathfrak{G}_{\mathrm{Prod}}(s) := \emptyset$;
  for all $m \in \mathrm{S}_{\mathrm{RNS}}$ mit $P(m) = s$ **do** *alle RNS, deren Produkt s ist*
    $\mathfrak{G}_{\mathrm{Prod}}(s) := \mathfrak{G}_{\mathrm{Prod}}(s) \cup \mathfrak{G}_{\mathrm{Prod}}(m)$;
  end for
else *s ist Dimer*
  Sei $s = \langle q, r \rangle$ mit $q, r \in \mathrm{S}_0$;
  $\mathfrak{G}_{\mathrm{Prod}}(s) := \mathfrak{G}_{\mathrm{Prod}}(q) \cup \mathfrak{G}_{\mathrm{Prod}}(r)$;
end if

Der Algorithmus 4.6 stellt den wichtigsten Teil bei der Erzeugung der Verbindungsstruktur dar: Durch die Wahl der Regulatoren für jede Produktionseinheit (PE) werden die Kanten zwischen den zugehörigen Genen im induzierten GRN-Graph festgelegt. Die Produktionseinheiten werden nacheinander erzeugt. Die Wahl, ob die nächste Produktionseinheit einen Selbstregulator erzeugen soll oder nicht, wird aufgrund der zur Gesamtzahl $\alpha_{\mathrm{self}} * N$ noch fehlenden Selbstregulatoren und der noch zu etablierenden Produktionseinheiten gefällt. Da hier nur die Verbindungsstruktur festgelegt wird, bleibt die Produktionsfunktion Prd noch unbestimmt.

Algorithmus 4.6 (Konstruktion einer Menge von Regulatoren für eine PE)
Eingabe: $s \in \mathrm{S}_0$ *zu produzierende Substanz*
k^m_{soll} mit $0 \leq k^m_{\mathrm{soll}} \leq k^m_{\mathrm{max}}$ *gewünschter Beitrag der PE zum multiplen Eingangsgrad*
$a \in \mathbb{B}$ *soll die PE einen Selbstregulator erzeugen?*
Ausgabe: Menge R von Regulatoren für die Produktion von s.

$k^m_{\mathrm{ist}} := 0$
$\mathrm{R} := \emptyset$
if a **then** *Selbstregulator*
  Wähle $G \in \mathfrak{G}_{\mathrm{Prod}}(s)$; *dieses Gen soll selbstregulierend werden*
  $\mathrm{S}_R := \{q \in \mathrm{S} : G \in \mathfrak{G}_{\mathrm{Prod}}(q)\}$; *alle von diesem Gen produzierten Substanzen*
  repeat
    $r := \mathrm{chooseSubstance}(k_{\mathrm{soll}}, \mathrm{S}_R)$; *Wähle geeignete Regulatorsubstanz*
  until geeignete Substanz gefunden?
  $\mathrm{R} := \mathrm{R} \cup \{r\}$
  $k^m_{\mathrm{ist}} := k^m_{\mathrm{ist}} + \mathrm{card}(\mathfrak{G}_{\mathrm{Prod}})$
end if
$\mathrm{S}_R := \mathrm{S}$; *mögliche Substanzen für Regulation*
if $\neg a$ **then**
  PE soll kein Selbstregulator sein, d.h. kein Gen, das s produziert darf einen Regulator für s produzieren:
  for all $G \in \mathfrak{G}_{\mathrm{Prod}}(s)$ **do**
    $\mathrm{S}_R := \mathrm{S}_R \setminus \{q \in \mathrm{S} : G \in \mathfrak{G}_{\mathrm{Prod}}(q)\}$;
  end for
end if

while $k_{\text{ist}} < k_{\text{soll}}$ **do**
  $r :=$ chooseSubstance$(k_{\text{soll}} - k_{\text{ist}}, \mathrm{S}_R)$;
  if geeignete Substanz gefunden? **then**
    $\mathrm{R} := \mathrm{R} \cup \{r\}$
    $k^m_{\text{ist}} := k^m_{\text{ist}} + \mathrm{card}(\mathfrak{G}_{\text{Prod}})$
  end if
end while
for all $G \in \mathfrak{G}_{\text{Prod}}(s)$ **do** *Update der multiplen Eingangsgrade aller Produzenten von s*
  $k^m_G := k^m_G + k^m_{\text{ist}}$
end for

Dieses Verfahren benutzt die Hilfsfunktion chooseSubstance(), die eine Substanz aus einer vorgegebenen Menge S_R von Substanzen auswählt, die maximal k Produzenten besitzt. Dabei werden die von *Hub*-Genen produzierten Substanzen bevorzugt. Es kann so allerdings nicht garantiert werden, dass andere Gene keinen ebenso hohen Ausgangsgrad wie die *Hubs* erhalten. Insbesondere Gene, die eine große Anzahl von Produkten aufweisen, neigen bei diesem Vorgehen zu einem ebenso hohen Ausgangsgrad wie die *Hubs*.

Algorithmus 4.7 (Hilfsfunktion zur Auswahl einer geeigneten Regulatorsubstanz)
function $r =$ chooseSubstance(k, S_R)
  Wähle $e \in [0, 1]$
  if $e \leq \alpha_{\text{ext}}$ **then** *externer Regulator*
    Wähle $r \in \mathrm{S}_R$ mit $\mathrm{card}(\mathfrak{G}_{\text{Prod}}(r)) = 0$;
  else
    Wähle $r \in \mathrm{S}_R$ mit $0 < \mathrm{card}(\mathfrak{G}_{\text{Prod}}(r)) \leq k$, bevorzuge dabei Substanzen, die auch in H sind;
  end if
end function

Das Verfahren kann so modifiziert werden, dass anstelle der Verteilung des multiplen Eingangsgrads die Verteilung des einfachen Eingangsgrads vorgegeben werden kann: Im Algorithmus 4.6 muss dafür eine Menge $\mathfrak{G}_R$ von regulierenden Genen geführt werden, deren Größe dem einfachen Eingangsgrad entspricht. Diese Menge ist am Anfang leer und wird mit jedem neuen Regulator $r \in \mathrm{S}$ erweitert: $\mathfrak{G}_R := \mathfrak{G}_R \cup \{G \in \mathfrak{G} : G \in \mathfrak{G}_{\text{Prod}}(r)\}$. Außerdem muss die Funktion chooseSubstance() entsprechend modifiziert werden.

4.2.2. Dynamisches Verhalten

Bei dem soweit erzeugten künstlichen Genregulationssystem sind noch die Zerfallsraten, die externen Änderungsraten und die Produktionsfunktionen zu bestimmen. Dafür gibt es verschiedene Möglichkeiten, insbesondere für die kinetische Beschreibung der Genexpression existieren eine Reihe von Ansätzen (z.B. Mendes, 2000; Mestl et al., 1995; D'haeseleer, 2000). Für erste Untersuchungen wurde folgende konkrete Ausprägung der Modellierung gewählt, wobei die zur Auswahl verwendeten Intervalle in Tabelle 4.3 angeführt sind:

Für jede Substanz $s \in \mathrm{S}$ wird die Zerfallskonstante δ_s gleichverteilt aus dem Intervall I_δ gewählt. Die externe Änderung y_s wird für jeden externen Regulator $s \in \mathrm{S}$ auf einen gleichverteilt aus I_y gewählten, konstanten Wert gesetzt. Für alle anderen Substanzen $q \in \mathrm{S}$ sei die externe Änderung konstant Null: $y_q(t) \equiv 0$.

Die Produktfunktionen werden, dem Modell von D'haeseleer (2000) folgend, als Sigmoidfunktion der gewichteten Summe der Eingänge gewählt: Jeder Produktionseinheit $\Pi = (s, \mathrm{R}, \mathrm{Prd})$ mit $\mathrm{R} = \{r_1, \ldots, r_l\}$ wird ein Vektor reeller Zahlen $(w_1, \ldots, w_l)$ (Gewich-

te) zugeordnet. Die Gewichte werden normalverteilt um den Wert 0 mit Standardabweichung 1 gewählt. Außerdem hat jede Einheit eine maximale Änderungsrate $d_{\max}$ und eine Grundaktivierung b, die gleichverteilt aus I_d bzw. I_b gewählt werden. Damit ergibt sich für die Transkription ($s \in \mathrm{S_{RNS}}$) folgende Gleichung:

$$\mathrm{Prd}(x_{r_1}(t), \ldots, x_{r_l}(t)) \quad =_{df} \quad d_{max} \times \mathrm{sig}\Big(b + \sum_{i=1}^{l} w_i \times x_i(t)\Big) \tag{4.39}$$

und für die Translation ($s = P(m), m \in \mathrm{S_{RNS}}$):

$$\mathrm{Prd}(x_{r_1}(t), \ldots, x_{r_l}(t), x_m(t)) \quad =_{df} \quad d_{max} \times x_m(t) \times \mathrm{sig}\Big(b + \sum_{i=1}^{l} w_i \times x_i(t)\Big), \tag{4.40}$$

wobei als Sigmoidfunktion

$$\mathrm{sig}(u) \quad =_{df} \quad \frac{1}{1 + e^{-u}} \tag{4.41}$$

verwendet wird. Damit wird angenommen, dass bei der Translation die Konzentration $x_m(t)$ des Primärtranskripts $m \in \mathrm{S_{RNS}}$ multiplikativ eingeht und dass sich die Einflüsse der Regulatoren additiv zusammensetzen.

Für eine konkrete Implementierung des hier beschriebenen Verfahrens ist es notwendig, die Intervalle festzulegen, aus denen die kinetischen Parameter und die Anfangskonzentrationen gleichverteilt gewählt werden. Die Tabelle 4.3 enthält die kinetischen Parameter und die Bezeichner für die Intervalle. Außerdem sind die Werte angegeben, die in der vorliegenden Implementierung (Abschnitt 4.3) verwendet werden und die auch den Experimenten in Kapitel 5 zugrunde liegen. Da keine biologischen Daten zur Abschätzung dieser Parameter vorlagen, wurden die Intervalle willkürlich gewählt.

Tab. 4.3.: Intervalle, aus denen die kinetischen Parameter gleichverteilt gewählt werden.

kinetischer Parameter	Variable	Intervall
Neigung zur Zusammenlagerung	I_γ	$[0.01, 0.21)$
Zerfallskonstante	I_δ	$[0.1, 0.4)$
externe Änderung	I_y	$[0, 2)$
maximale Änderungsrate	I_d	$[0.1, 1.0)$
Grundaktivierung	I_b	$[0, 0.01)$
Anfangskonzentration	I_0	$[0, 2)$

4.3. Implementierung

Die in Abschnitt 4.1 beschriebene Modellierung wurde in das `Java`-Paket `artGED` umgesetzt. Dabei finden sich die meisten definierten Begriffe als entsprechende Klassen wieder. Die Hauptklasse des Projektes `artGED.GeneRegulatorySystem` realisiert ein Genregulationssystem. Durch den Konstruktor ist es möglich, ein künstliches Genregulationssystem entsprechend der Vorgaben einer Parameterdatei zu erzeugen. Dafür wird das im letzten Abschnitt erläuterte Verfahren benutzt. Es erfolgt sowohl eine Ausgabe des multiplen als auch des einfachen induzierten GRN-Graphen. Außerdem wird ein entsprechendes Differenzialgleichungssystem ausgegeben. Die Ein- und Ausgaben des Programms sind in

Abbildung 4.7 veranschaulicht. Durch eine ebenfalls ausgegebene Parameterdatei, die eine Identifikationsnummer des erzeugten künstlichen Genregulationssystems enthält, ist es möglich, dieses Genregulationssystem zu einem späteren Zeitpunkt erneut zu erzeugen. Dies gestattet verschiedenen Nutzern über ein- und dasselbe Genregulationssystem zu verfügen.

Zum derzeitigen Entwicklungsstand ist die Simulation des dynamischen Verhaltens eines Genregulationssystems nicht in das `Java`-Programm integriert. Statt dessen erfolgt eine Ausgabe des Differenzialgleichungssystems als `Octave`-Quelldatei. Mit Hilfe des in `Octave` verfügbaren Differenzialgleichungslösers kann eine Lösung für das Differenzialgleichungssystem gesucht werden. So können Experimente simuliert und künstliche Genexpressionsdaten erhalten werden.

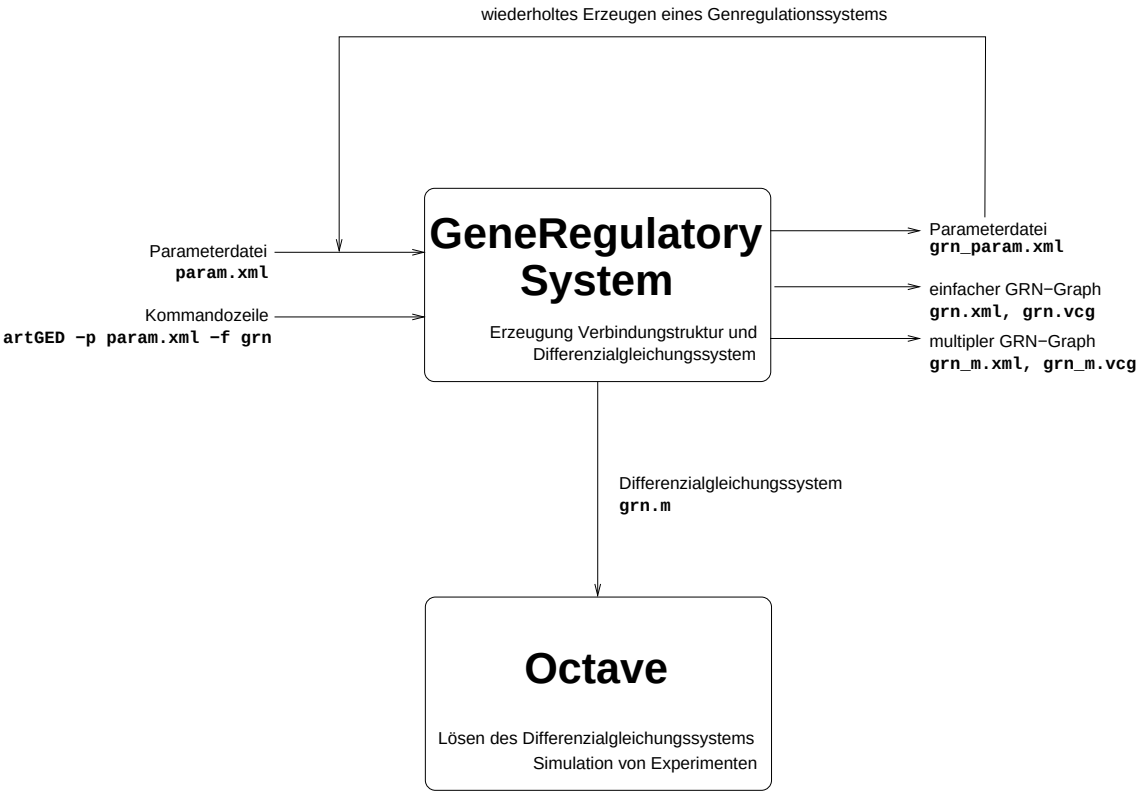

Abb. 4.7.: Ablauf bei der Erzeugung künstlicher Genexpressionsdaten und Ein-/Ausgabe der Programmteile (beispielhaft mit `Dateinamen` versehen).

Der erste Schritt, die Erzeugung eines Genregulationssystems, benötigt wesentlich weniger Rechenressourcen als die numerische Lösung des Differenzialgleichungssystems. So benötigte die Erzeugung eines Systems mit 1000 Genen (die Parameterwahl entsprach Tabelle 5.4) auf einem `LINUX`-System mit Prozessor `AMD Athlon(tm) XP 2400+` und 512 MB Arbeitsspeicher etwa 2 Sekunden. Für die Lösung des entsprechenden Differenzialgleichungssystems, bestehend aus 1885 Gleichungen, benötigte `Octave` drei Stunden und 28 Minuten. Für die Eingabe der Parameter und die Ausgabe des multiplen und des einfachen GRN-Graphen wurde ein `XML`-basiertes (Bray et al., 2000) Dateiformat verwendet. In Anhang A sind die entsprechenden `XML`-Schemata angeführt. Die GRN-Graphen werden zusätzlich in einem Format ausgegeben, das von dem Visualisierungswerkzeug `XVCG` von Sander (1995) gelesen werden kann. Damit können die GRN-Graphen grafisch dargestellt werden.

Durch die Form der verwendeten Produktfunktionen (gewichtete Summe der Eingänge) kann den im GRN-Graph dargestellten Regulationseinflüssen eine Stärke zugeordnet werden: Im multiplen GRN-Graphen stellt jede Kante einen regulatorischen Einfluss dar, dessen Stärke dem Gewicht in der zugehörigen Differenzialgleichung entspricht. Dagegen können Kanten im einfachen GRN-Graphen mehrere Regulationseinflüsse repräsentieren, deren einzelne Stärken sich additiv zu einer Gesamtstärke zusammensetzten. Die Stärke einer Kante wird in der `XML`-Datei als numerischer Wert und in der `XVCG`-Datei als Kantenattribut („inhibierend“ bei negativer Stärke, sonst „aktivierend“) angegeben.

Für die Eingangsgradverteilung K^m sind zur Zeit zwei mögliche Vorgaben implementiert: Gleichverteilung zwischen 0 und $k^m_{\max}$ ($0 < k^m_{\max} < N$) und Normalverteilung mit vorgebbarem Mittelwert $0 < \overline{k^m} < N$ und Standardabweichung σ. Bei der Normalverteilung wird ebenfalls ein maximaler Eingangsgrad gesetzt, da der Algorithmus dadurch stabiler wird:

$$k^m_{\max} = 2 \times (\overline{k^m} + \sigma), \tag{4.42}$$

jedoch nicht größer als $N-1$. Falls bei Normalverteilung ein Wert außerhalb des zulässigen Bereichs $\{0, \ldots, k^m_{\max}\}$ gezogen wird, erfolgt solange ein erneutes Ziehen, bis ein gültiger Wert gefunden wurde.

Das vorliegende Programm ist in der Lage, künstliche Genregulationssysteme zu erzeugen. In Zusammenarbeit mit dem mathematischen Programmpaket `Octave` (Eaton, 1998) können aus den zugehörigen Differenzialgleichungssystemen Genexpressionsdaten gewonnen werden. Da für dieses Zusammenspiel noch keine Nutzerschnittstelle geschaffen wurde, müssen Experimente zur Messung von Genexpressionsdaten „von Hand“ formuliert werden. Die in Kapitel 5.3 beschriebenen experimentellen Untersuchungen zum dynamischen Verhalten künstlicher Genregulationssysteme wurden auf diese Weise durchgeführt. Im Folgenden wird dieses Vorgehen skizziert.

Die ausgegebene `Octave`-Quelldatei enthält:

1. das Differenzialgleichungssystem als `function xdot=grsode(x,t)`,
2. symbolische Konstanten, über die die Substanzen referenziert werden können,
3. für jede Substanz $s \in \mathrm{S}$ eine aus I_0 (siehe Tabelle 4.3) gleichverteilt gewählte Anfangskonzentration `x0(s)` und
4. für jede Substanz $s \in \mathrm{S}$ zwei Variablen `ext(s)=0.0` und `a(s)=1.0` (siehe unten).

Eine Differenzialgleichung hat dabei folgende Form (am Beispiel für eine Substanz $m_5 \in \mathrm{S_{RNS}}$):

```
xdot(m5) = ext(m5)+ a(m5) * (0.286 * g(0.0090+ 0.657 * x(P4))) - 0.215 * x(m5);
```

Dabei entspricht `xdot(m5)` dem Differenzialquotienten $\frac{dx_{m_5}}{dt}$, `ext(m5)` ist die externe Änderung y_{m_5}, der letzte Term ist der Zerfall mit $\delta_{m_5} = 0.215$ und der mittlere Term die Transkriptionsfunktion mit dem Regulator $P_4 \in \mathrm{S_{Prot}}$. g stellt die oben besprochene Sigmoidfunktion dar. Die Teile der Differenzialgleichung einer Substanz $s \in \mathrm{S}$, die für erzeugende Prozesse stehen (Produktion bei einfachen Substanzen und Bildung bei Dimeren), sind außerdem mit einem Faktor `a(s)` versehen, der auf den Wert 1 voreingestellt ist.

Durch Lösen des Differenzialgleichungssystems können künstliche Genexpressionsdaten (Matrix bzw. Vektor `x` von Messwerten) gewonnen werden:

1. Zeitserien, z.B. zwischen dem Zeitpunkt $t = 0$ und $t = 100$ mit 300 Messpunkten:

   ```
   grn;
   t=linspace(0,100,300);
   x=lsode("grsode",x0,t);
   ```

2. Gleichgewichtsmessung, z.B. zu einem Zeitpunkt $t = 300$:

   ```
   grn;
   x=lsode("grsode",x0,[300]);
   ```

Im jeweils ersten Schritt werden die symbolischen Konstanten, die Variablen und das Differenzialgleichungssystem in der Quelldatei `grn.m` dem `Octave`-System bekannt gemacht.

Mit den Variablen `ext(s)` und `a(s)` können leicht beliebige Störexperimente in `Octave` formuliert werden, z.B.:

1. die Löschung eines Gens $m_3 \in \mathrm{S_{RNS}}$:
   `ext(m3)=0; a(m3)=0;`
2. die doppelte Überexprimierung von $m_9 \in \mathrm{S_{RNS}}$
   `a(m9)=2;`
3. die zeitlich begrenzte Störung durch Zugabe eines Proteins $P_{13} \in \mathrm{S}$:

   ```
   function f(t)
     if t<10
      0.7;
     else
      0.0;
     end
   endfunction

   ext(P13)=f(t);.
   ```

Nachdem ein Störexperiment auf diese Weise festgelegt wurde, können Genexpressionsdaten auf eine der beiden oben beschriebenen Arten (Zeitserie oder Gleichgewichtsmessung) erzeugt werden.

5. Experimentelle Untersuchungen

Mit dem im letzten Kapitel beschriebenen Verfahren wurden Versuche durchgeführt, um die Leistungsfähigkeit zu überprüfen und Eigenschaften der erzeugten Genregulationssysteme zu untersuchen. Zuerst werden die GRN-Graphen kleinerer Netze dargestellt, um einen Eindruck von den erzeugten Systemen zu vermitteln. Anschließend werden Eigenschaften der Verbindungsstruktur untersucht. Im letzten Abschnitt werden Experimente zum dynamischen Verhalten erzeugter Differenzialgleichungssysteme durchgeführt.

Für die hier beschriebenen Versuche wurden die in Tabelle 5.4 aufgeführten Parameter als Standard genommen. Im Folgenden werden nur die jeweils veränderten Parameter genannt. Bei den meisten Versuchen sind die gezeigten Werte über $n_{\mathrm{GRS}} = 1000$ unabhängige Programmläufe mit identischen Parametern gemittelt.

Tab. 5.4.: Standardparameter für die durchgeführten Versuche.

Parameter	**Symbol**	**Standardwert**
Größe	N	1000
Anzahl RNS	n_{rns}	$1.2 \times N$
Anzahl Proteine	n_{prot}	$1.2 \times N$
Anteil nicht kodierender RNS	$\alpha_{\perp}$	0.05
Anzahl Homodimere	n_{hom}	$0.3 \times N$
Anzahl Heterodimere	n_{het}	$0.2 \times N$
Dimerisationsneigung von RNS	γ_{rns}	1.0
Anzahl *Hubs*	n_{hub}	2
Anteil Selbstregulatoren	α_{self}	0.02
Anteil externer Regulatoren	α_{ext}	0.0
maximaler multipler Eingangsgrad	k^m_{max}	13
Eingangsgradverteilung	K^m	Normalverteilung: $\overline{k^m} = 4, \sigma = 2.5$
Regulationsneigung von RNS	ρ_{rns}	1.0
Anteil der transkriptionalen Regulation	α_{tkr}	0.5
Anzahl erzeugter Genregulationssysteme	n_{GRS}	1000

5.1. Erzeugung von Genregulationssystemen

Um einen Eindruck von den erzeugten GRN-Graphen zu gewinnen, werden in diesem Abschnitt kleine Genregulationssysteme vorgestellt, die mit dem oben beschriebenen Verfahren erzeugt wurden. Hierbei wurde jeweils nur ein einziges Genregulationssystem erzeugt ($n_{\mathrm{GRS}} = 1$).

5.1.1. Vergleich einfacher und multipler GRN-Graph

Die Abbildung 5.8 stellt den multiplen und den einfachen GRN-Graphen eines Genregulationssystems mit 20 Genen gegenüber. Damit die GRN-Graphen übersichtlich bleiben, wurde ein mittlerer multipler Eingangsgrad $\overline{k^m} = 2$ gewählt. Deutlich zu erkennen ist, dass der einfache GRN-Graph wesentlich weniger Information als der multiple GRN-Graph enthält.

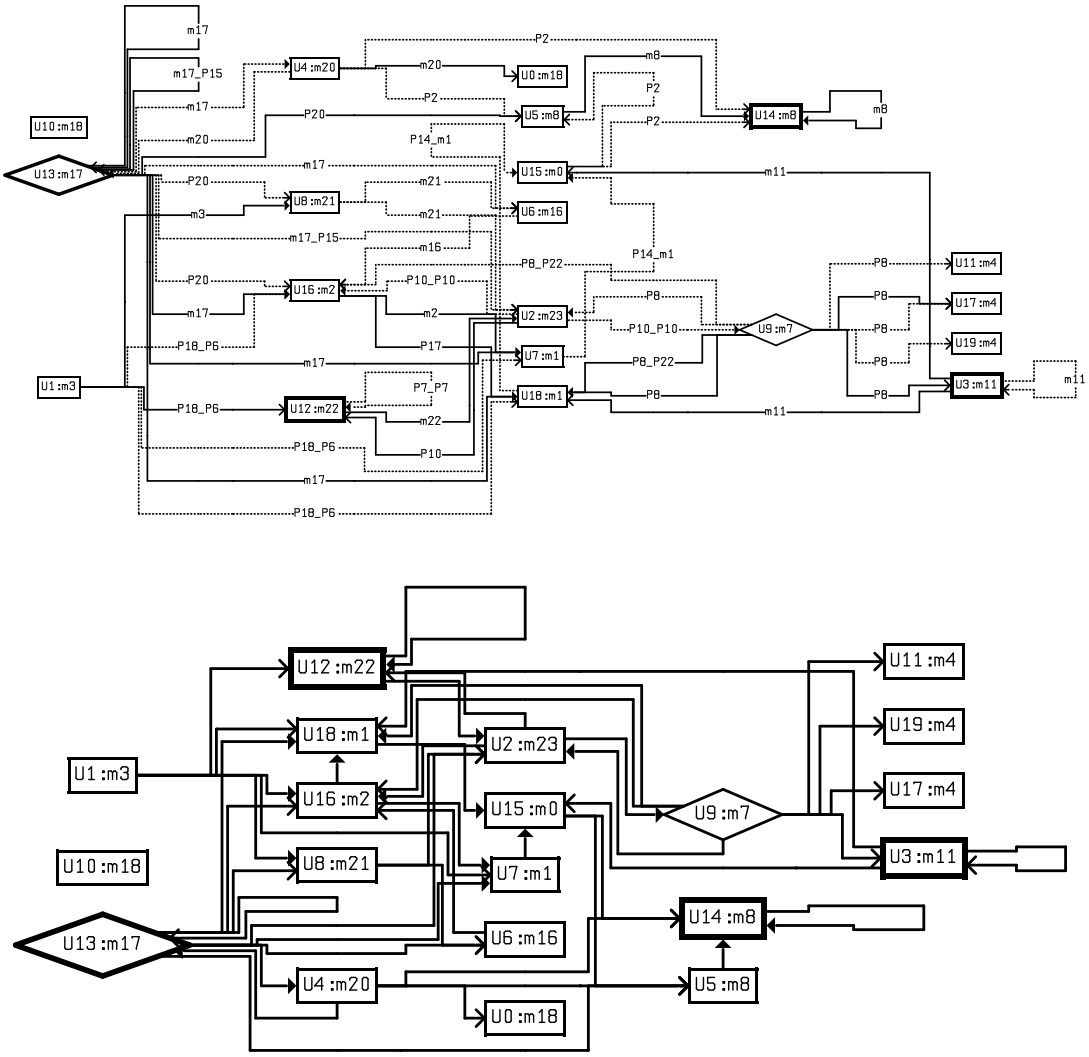

Abb. 5.8.: Multipler (oben) und einfacher (unten) GRN-Graph mit $N = 20$ Genen. Bedeutung der Symbole: viereckiger Rahmen = Gen mit Name und Primärtranskript, dicker Rahmen = Selbstregulator, Rhombus = *Hub*, durchgezogene Linie = transkriptionale Regulation, gepunktete Linie = posttranskriptionale Regulation, Linienbeschriftung = regulierende Substanz, volle Pfeilspitze = aktivierender Einfluss, offene Pfeilspitze = inhibierender Einfluss, Kreis = externer Regulator. $k^m_{\max} = 7$, $\overline{k^m} = 2$, $\sigma = 1.5$, $n_{\text{GRS}} = 1$.

5.1.2. Regulationsmechanismus

Das hier vorgestellte Verfahren gestattet eine wesentlich detailliertere Modellierung der Genexpression als die meisten in Kapitel 3.2.1 vorgestellten künstlichen Modellsysteme. Es wird Regulation sowohl auf der transkriptionalen als auch der posttranskriptionalen Ebene berücksichtigt. Die Bildung von Dimeren wird explizit betrachtet. Regulation kann durch Proteine aber auch durch RNS und Dimere erfolgen.

Durch geeignete Wahl der Parameter ist es jedoch auch möglich, ein stark vereinfachtes Modellsystem zu erhalten (Abbildung 5.9). Der durch diese Parameterwahl erreichte Modellierungsgrad entspricht etwa dem der Modellsysteme, die von Kyoda et al. (2000) für die Erzeugung von Testdaten verwendet wurden. Demgegenüber stellt Abbildung 5.10 den multiplen GRN-Graphen eines reichhaltigen Modellsystems mit den oben angeführten Details dar.

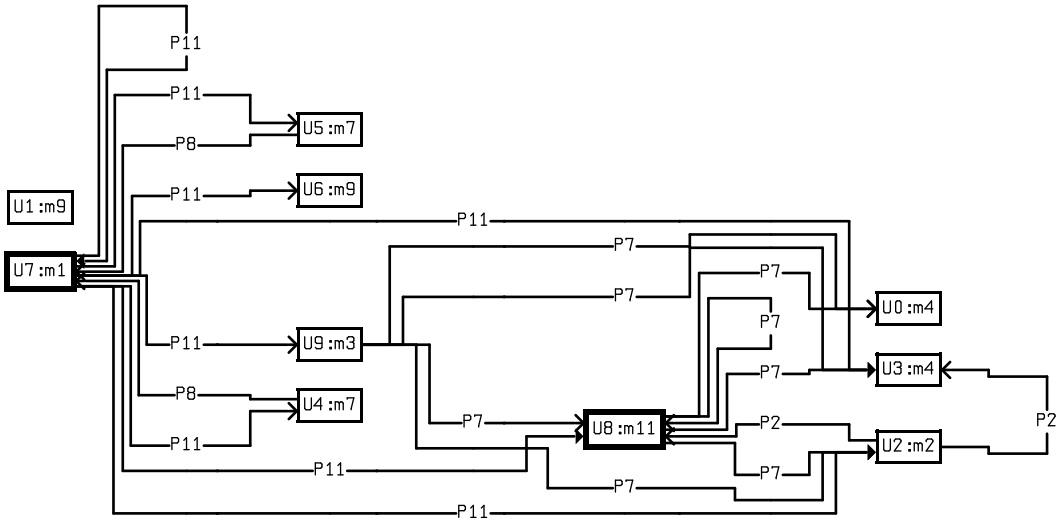

Abb. 5.9.: Multipler GRN-Graph eines einfachen Genregulationssystems: $N = 10$, $\alpha_{\perp} = 0.0$ (alle RNS sind kodierend), $n_{\text{hom}} = 0$, $n_{\text{het}} = 0$ (keine Dimere), $\gamma_{\text{rns}} = 0.0$, $n_{\text{hub}} = 0$ (keine *Hubs*), $k^m_{\max} = 7$, $\overline{k^m} = 2$, $\sigma = 1.5$, $\rho_{\text{rns}} = 0.0$ (RNS wirkt nicht regulierend), $\alpha_{\text{tkr}} = 1.0$ (nur transkriptionale Regulation), $\alpha_{\text{ext}} = 0.0$ (keine externen Regulatoren), $n_{\text{GRS}} = 1$. Symbole wie in Abbildung 5.8.

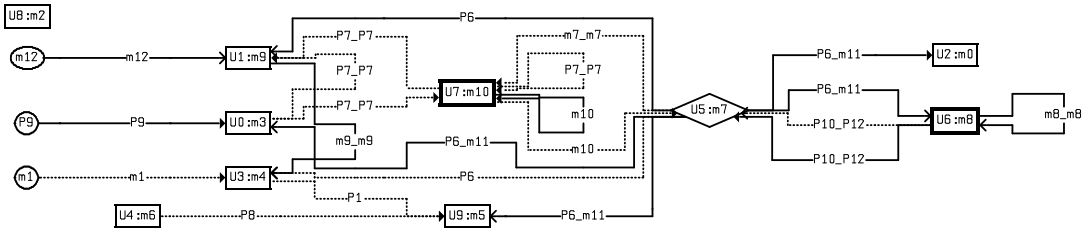

Abb. 5.10.: Multipler GRN-Graph eines reichhaltigen Genregulationssystems: $N = 10$, $n_{\text{hom}} = 8$, $n_{\text{het}} = 8$, $n_{\text{hub}} = 1$, $\alpha_{\text{ext}} = 0.1$, $k^m_{\max} = 7$, $\overline{k^m} = 2$, $\sigma = 1.5$, $n_{\text{GRS}} = 1$. Symbole wie in Abbildung 5.8.

5.2. Verbindungsstruktur

Dieser Abschnitt untersucht Eigenschaften der Verbindungstruktur der erzeugten Genregulationssysteme. Jeder gezeigte Wert ist der Mittelwert über $n_{\text{GRS}} = 1000$ unabhängige Programmläufe mit identischen Parametern.

5.2.1. Vergleich multipler und regulatorischer Eingangsgrad

Hier werden die Aussagen 3 und 4 von Satz 4.2 veranschaulicht. In Abbildung 5.11 wird gezeigt, dass sich der regulatorische Eingangsgrad mit wachsendem Anteil α_{ext} von externen Regulatoren dem multiplen Eingangsgrad annähert und diesen schließlich übersteigt. Ohne externe Regulatoren liegt der regulatorische Eingangsgrad deutlich unter dem multiplen Eingangsgrad. Das Verfahren sorgt für eine Einhaltung der vorgegebenen Verteilung des multiplen Eingangsgrads. Mit steigendem Anteil an externen Regulatoren, werden diese immer häufiger von chooseSimpleSubstance() (Algorithmus 4.7) ausgewählt. Diese tragen aber nur zu einer Erhöhung des regulatorischen, nicht aber des multiplen Eingangsgrads bei.

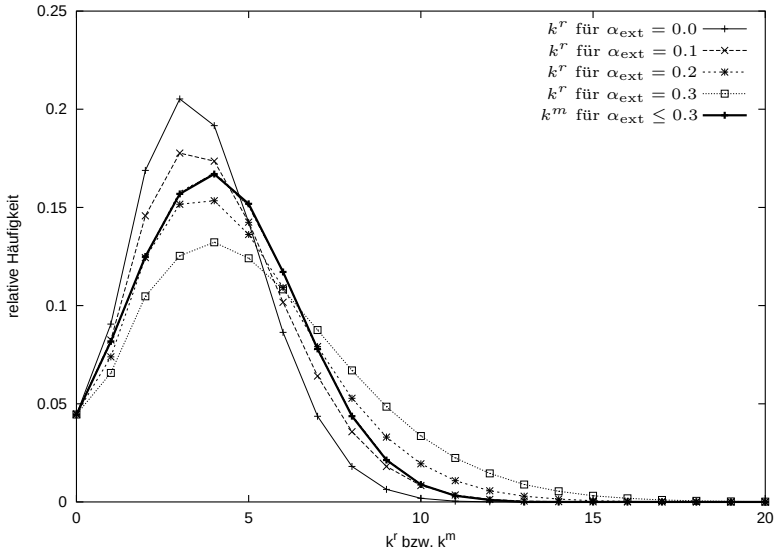

Abb. 5.11.: Relative Häufigkeiten von Genen mit bestimmtem regulatorischem (k^r) bzw. multiplem (k^m) Eingangsgrad bei unterschiedlichem Anteil α_{ext} externer Regulatoren. Die Graphen für k^m für alle vier Versuche fallen zusammen. Jeder Messpunkt ist Mittelwert aus $n_{\text{GRS}} = 1000$ unabhängigen Programmläufen.

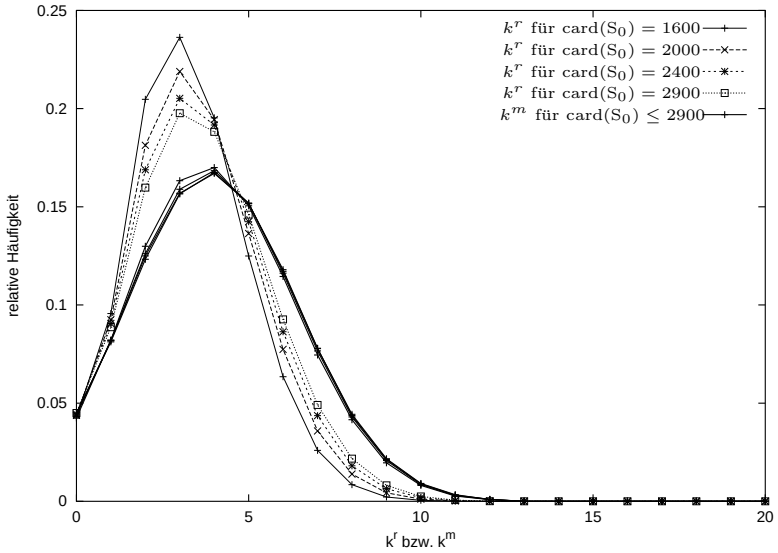

Abb. 5.12.: Relative Häufigkeiten von Genen mit bestimmtem regulatorischem (k^r) bzw. multiplem (k^m) Eingangsgrad bei unterschiedlicher Anzahl card(S_0) von einfachen Substanzen. Dabei gilt: $n_{\text{rns}} = n_{\text{prot}} = \frac{1}{2}\,\text{card}(S_0)$. Die Graphen für k^m sind annähernd gleich. Jeder Messpunkt ist Mittelwert aus $n_{\text{GRS}} = 1000$ unabhängigen Programmläufen.

Die Anzahl der Produzenten eines Regulators hat, wie Satz 4.2 zeigt, ebenfalls Einfluss auf den Unterschied zwischen multiplem und regulatorischem Eingangsgrad: Je mehr Produzenten ein Regulator besitzt, umso größer ist sein Beitrag zum multiplen Eingangsgrad. Der regulatorische Eingangsgrad wird von jedem Regulator jedoch nur um den Wert 1 erhöht. Im oben beschriebenen Verfahren wird durch die Wahl der Produktfunktion und der Primärtranskripte (Algorithmen 4.1 und 4.4) die Anzahl der Produzenten einer Substanz festgelegt: Besitzen mehrere Gene dasselbe Primärtranskript, erhöht sich die Anzahl der Produzenten von diesem und daraus abgeleiteten Substanzen (translatiertes Protein und Dimere). Ebenso kann die Anzahl von Produzenten eines Proteins (und der Dimere, an denen es beteiligt ist) dadurch erhöht werden, dass mehrere RNS

dieses als Produkt besitzen. Die durchschnittliche Anzahl von Produzenten wird in dem beschriebenen Verfahren also stark von der Anzahl card(S_0) einfacher Substanzen beeinflusst: Eine höhere Anzahl lässt mehr Möglichkeiten zu, ein Primärtranskript (Algorithmus 4.4) bzw. ein Produkt (Algorithmus 4.1) zu wählen und führt somit zu einer geringeren durchschnittlichen Anzahl von Produzenten. In Folge bedeutet dies eine Annährung des regulatorischen an den multiplen Eingangsgrad. Diese Abhängigkeit zeigt Abbildung 5.12.

5.2.2. Variation der Eingangsgradverteilung

In diesem Abschnitt wird untersucht, inwieweit das Verfahren in der Lage ist, Vorgaben in der Verteilung des multiplen Eingangsgrads zu realisieren. Dafür wurden für verschiedene vorgegebene Gleich- und Normalverteilungen des multiplen Eingangsgrads jeweils $n_{\mathrm{GRS}} = 1000$ Genregulationssysteme erzeugt. Für diese wurden die relativen Häufigkeiten von Genen mit einem bestimmten Eingangsgrad gemessen. Die Abbildung 5.13 zeigt deutlich, dass die Vorgaben im betrachteten Bereich sehr gut erfüllt werden.

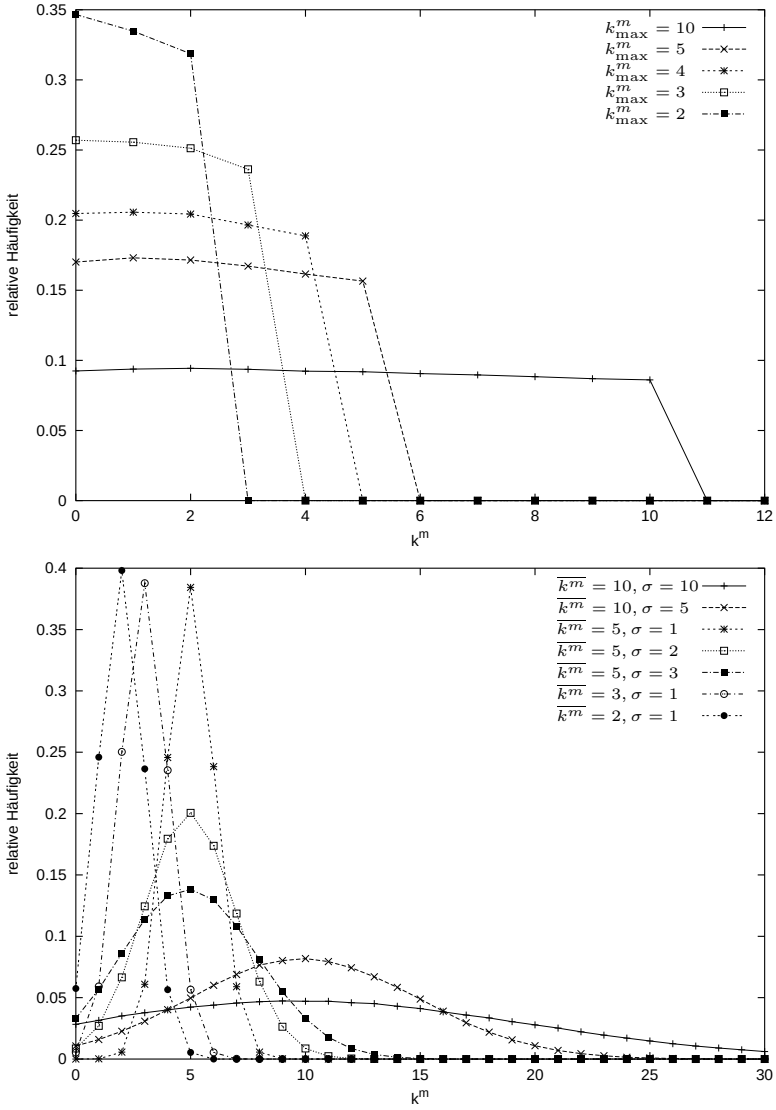

Abb. 5.13.: Relative Häufigkeiten von Genen mit bestimmtem multiplem Eingangsgrad k^m. Vorgegeben wurde: (oben) Gleichverteilung des multiplen Eingangsgrads zwischen 0 und $k^m_{\max}$, (unten) Normalverteilung um $\overline{k^m}$ mit Standardabweichung σ ($k^m_{\max} = 2(\overline{k^m} + \sigma)$). Jeder Messpunkt ist Mittelwert aus $n_{\mathrm{GRS}} = 1000$ unabhängigen Programmläufen.

5.2.3. Selbstregulatoren

Die vorgegebene Anzahl Selbstregulatoren kann vom Algorithmus 4.6 nicht immer erreicht werden (siehe Seite 51). Dafür gibt es zwei mögliche Ursachen: Zum einen gibt es Fälle, in denen ein gleichzeitiges Einhalten der geforderten Eingangsgradverteilung und der Anzahl der Selbstregulatoren prinzipiell unmöglich ist. Zum anderen ist dies durch den implementierten Algorithmus bedingt: Die Priorität des Verfahrens liegt im Erreichen des vorgegebenen multiplen Eingangsgrades. Bei der Auswahl einer Regulatorsubstanz zur Etablierung eines Selbstregulators (Algorithmus 4.6) kann es zu Fehlversuchen kommen, wenn die gewählte Substanz eine zu große Anzahl Produzenten besitzt. Nach einer bestimmten Anzahl solcher Fehlversuche bricht das Verfahren aus Effizienzgründen ab.

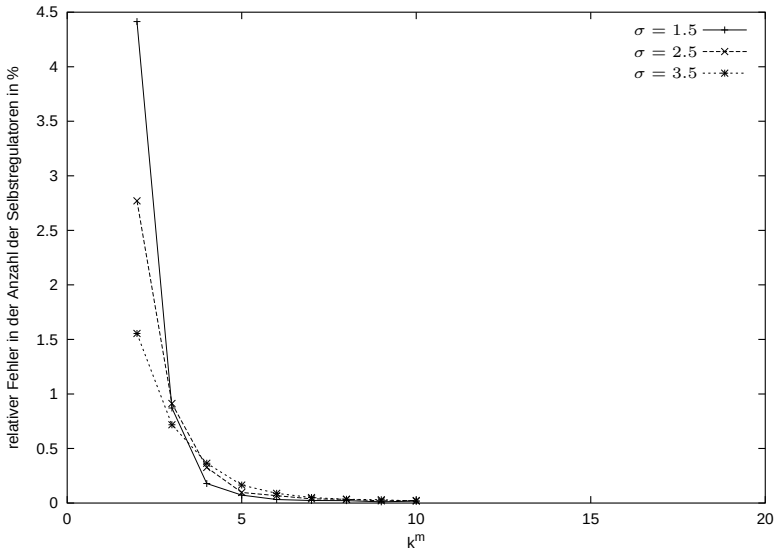

Abb. 5.14.: Relativer Fehler in der Anzahl der Selbstregulatoren in Abhängigkeit des mittleren multiplen Eingangsgrads $\overline{k^m}$ bei Normalverteilung mit unterschiedlichen Standardabweichungen σ ($k^m_{\max} = 2(\overline{k^m} + \sigma)$). Jeder Messpunkt ist Mittelwert aus $n_{\mathrm{GRS}} = 1000$ unabhängigen Programmläufen.

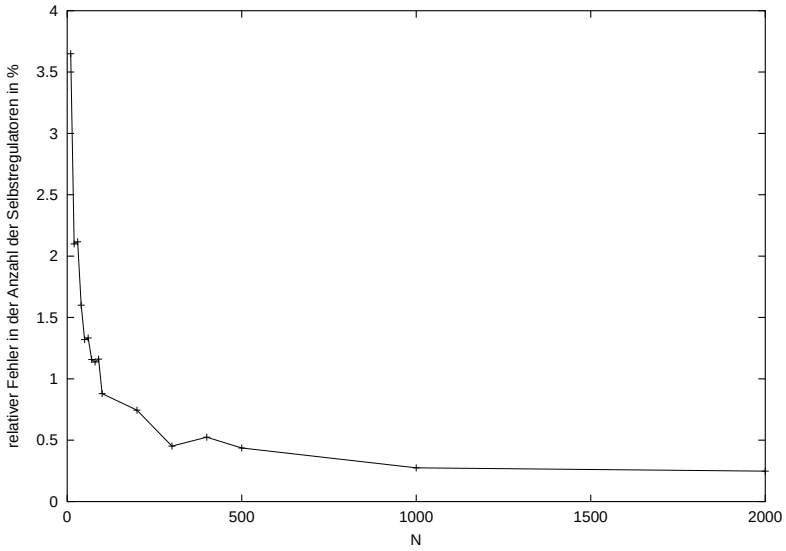

Abb. 5.15.: Relativer Fehler in der Anzahl der Selbstregulatoren in Abhängigkeit von der Netzgröße N. Jeder Messpunkt ist Mittelwert aus $n_{\mathrm{GRS}} = 1000$ unabhängigen Programmläufen.

Dabei hängt die erreichbare Anzahl von Selbstregulatoren entscheidend von der Größe des Genregulationssystems und von der Verteilung des multiplen Eingangsgrads ab: Je kleiner das Netz bzw. je schmaler die Verteilung des multiplen Eingangsgrads, umso weniger Möglichkeiten bei der Wahl einer geeigneten Regulatorsubstanz hat der Algorithmus 4.6. Dies führt zu einem Anwachsen des Fehlers in der Anzahl der Selbstregulatoren (Abbildungen 5.14 und 5.15). Wie zu sehen ist, liegt selbst bei ungünstigen Verhältnissen der relative Fehler in der Anzahl der Selbstregulatoren unter 5%.

5.3. Simulationen der Dynamik

Um einen Eindruck von der Dynamik eines Genregulationssystems zu vermitteln, sind in Abbildung 5.16 die zeitlichen Konzentrationsverläufe eines Systems mit $N = 20$ Genen dargestellt. Dabei wurden nur diejenigen Substanzen betrachtet, die entweder im System gebildet werden (durch Produktionseinheiten oder Dimerisation) oder einen Einfluss auf die Bildung anderer Substanzen ausüben. Wie die Abbildung veranschaulicht, streben die Konzentrationen aller Substanzen nach einer Anfangsphase ein Gleichgewicht an. In diesem sind die Bildungs- und die Abbaurate einer Substanz gleich groß, die Konzentration ändert sich nicht mehr. Dieses Verhalten wurde bei allen untersuchten Systemen beobachtet und ist in Übereinstimmung mit Stadler et al. (1993): Dort wurde gezeigt, dass zufällig erzeugte dynamische Systeme der hier verwendeten Art (gewichtete Summe, vergleiche Gleichungen 4.39 und 4.40) dazu neigen, nach einer Anfangsphase in einen stabilen Endzustand überzugehen.

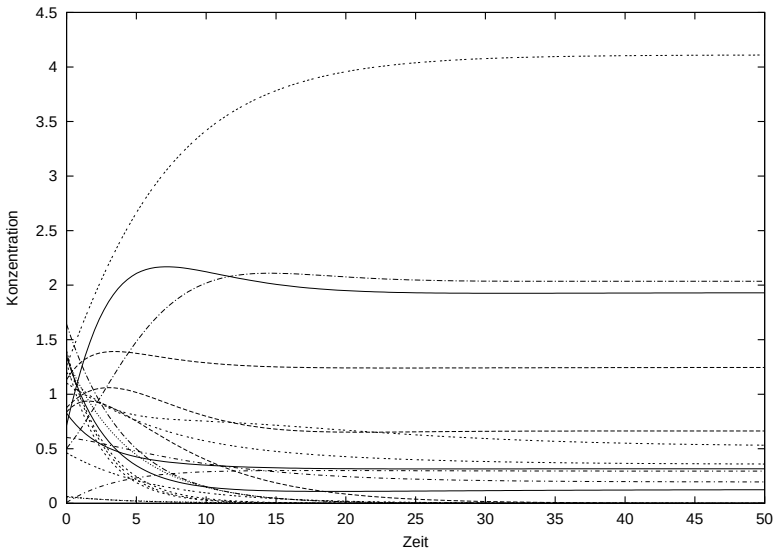

Abb. 5.16.: Darstellung der zeitlichen Konzentrationsverläufe von 20 Substanzen (13 RNS und 7 Proteine) für ein Genregulationssystem mit $N = 20$ Genen ($n_{\text{hom}} = 0$, $n_{\text{het}} = 0$, $k^m_{\text{max}} = 6$, $\overline{k^m} = 2$, $\sigma = 1.0$, $n_{\text{hub}} = 1$, $n_{\text{GRS}} = 1$).

5.3.1. Verteilung der Genaktivitäten

Im stationären Zustand ist die Konzentration nur weniger Substanzen hoch, der Großteil der Substanzen hat eine niedrige Konzentration. Vermutlich liegt dies zum einen daran, dass die Gewichte der Produktionsfunktionen (Gleichungen 4.39 und 4.40) normalverteilt um den Wert Null gewählt werden. Dadurch heben sich die meisten regulatorischen Einflüsse gegeneinander auf. Zum anderen kann die Wahl der kinetischen Parameter (vergleiche Tabelle 4.3) zu diesem Verhalten führen. Eventuell wird so der Zerfall zum dominierenden Prozess. Die Aufklärung der Ursachen dieses Phänomens und die Wahl geeigneter kinetischer Parameter sind einer weiteren Untersuchung wert.

Deutlich illustrieren die Histogramme der relativen Häufigkeiten der Substanzen des Genregulationssystems aus Abbildung 5.16 und eines sehr großen Systems mit $N = 1000$ Genen (1885 Substanzen) (Abbildungen 5.17 und 5.18) das genannte Verhalten.

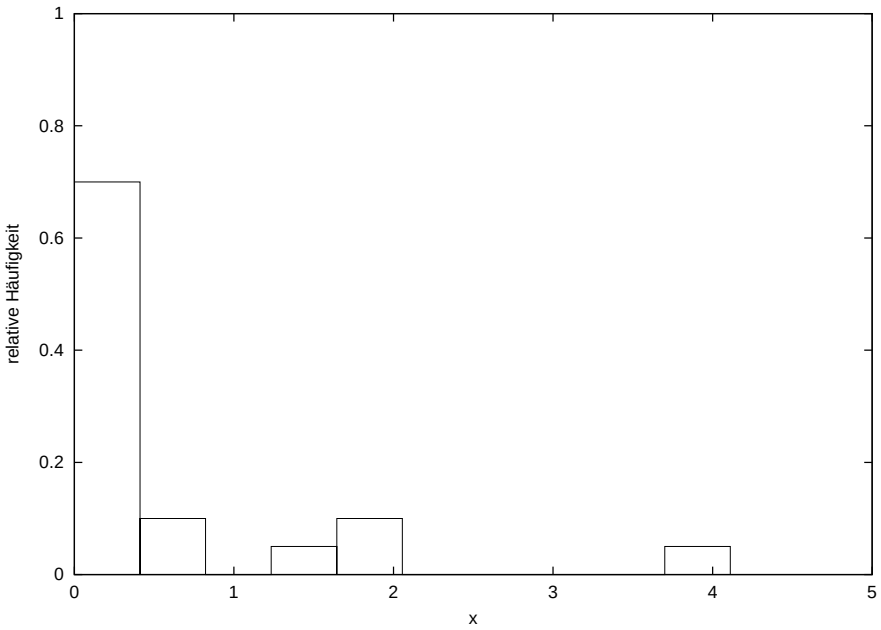

Abb. 5.17.: Histogramm der relativen Häufigkeiten von Substanzen mit einer bestimmten Konzentration des Genregulationssystems aus Abbildung 5.16 zum Zeitpunkt $t = 500$ mit insgesamt 20 Substanzen. Die maximale Konzentration zu diesem Zeitpunkt liegt bei 4.1.

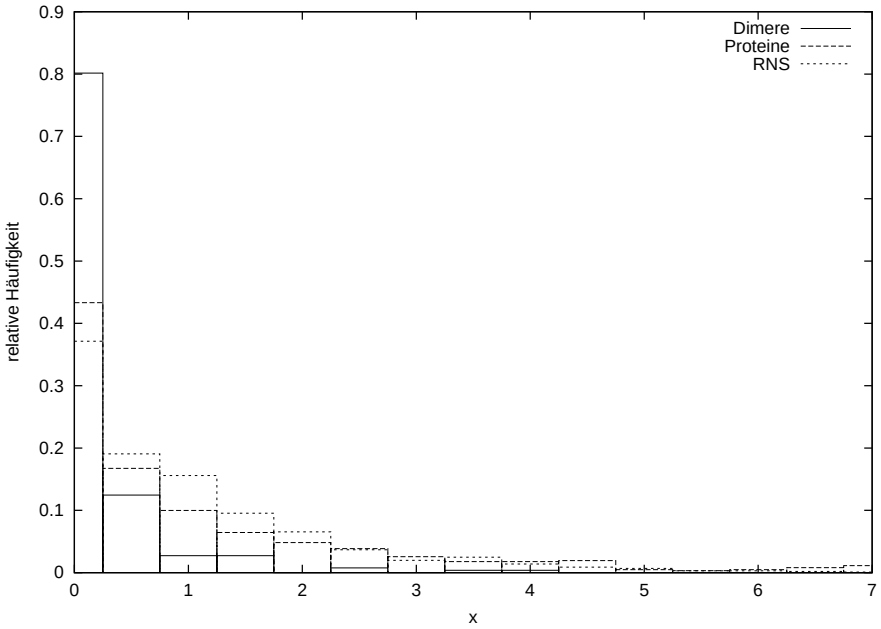

Abb. 5.18.: Histogramm der relativen Häufigkeiten von RNS, Proteinen bzw. Dimeren mit einer bestimmten Konzentration eines Genregulationssystems entsprechend den Standardparametern aus Tabelle 5.4 zum Zeitpunkt $t = 500$ mit insgesamt 1885 Substanzen ($n_{\mathrm{GRS}} = 1$). Die maximale Konzentration zu diesem Zeitpunkt liegt bei 22.1. Es sind nur die relativen Häufigkeiten von Konzentrationen unter 7 dargestellt, die relativen Häufigkeiten höherer Konzentrationen liegen unter 1%.

5.3.2. Aktivitätsänderungen bei einfachen Löschexperimenten

Aus dem Konzentrationshistogramm in Abbildung 5.18 ist ersichtlich, dass ein großer Teil der RNS im stationären Zustand des Systems eine sehr geringe Konzentration aufweist. Es ist daher zu erwarten, dass die Löschung eines beliebigen Gens im Allgemeinen nur eine recht kleine Änderung im dynamischen Verhalten des Gesamtsystems hervorruft. Weitere Faktoren, die ein Genregulationssystem unempfindlicher gegen Löschung einzelner Gene machen, sind Rückkopplungen und redundante Pfade, die in zufällig erzeugten, komplexen Verbindungsstrukturen zahlreich vorkommen.

Bei dem im Folgenden besprochenen Vorgehen werden nicht Gene sondern RNS „gelöscht“. Da mehrere Gene dasselbe Primärtranskript besitzen können und nicht alle im Genregulationssystem vorkommenden Substanzen auch in das Differenzialgleichungssystem übernommen werden (siehe oben), variiert die Anzahl l der Experimente. Im Durchschnitt wurden $l = 95$ Löschexperimente pro Genregulationssystem durchgeführt. Sei

$$\mathrm{S}^*_{\mathrm{RNS}} =_{df} \{m_1, \ldots, m_l\} \subseteq \mathrm{S}_{\mathrm{RNS}} \tag{5.1}$$

die Menge der im Differenzialgleichungssystem vorkommenden RNS.

Für Genregulationssysteme mit $N = 100$ Genen wurden jeweils alle einfachen RNS-Löschexperimente in folgender Weise simuliert:

1. Das ungestörte Differenzialgleichungssystem wurde für einen festgelegten Zeitpunkt $t = 200$ gelöst. Es wird angenommen, dass zu diesem Zeitpunkt das System seinen stationären Zustand erreicht hat. Der so erhaltene Konzentrationsvektor des Wildtyps wird mit $\mathbf{x}^{\mathrm{wt}}(t) =_{df} (x_1^{\mathrm{wt}}(t), \ldots, x_n^{\mathrm{wt}}(t))$ bezeichnet, wobei n die Anzahl der Differenzialgleichungen sei.
2. Für jede im Differenzialgleichungssystem vorkommende RNS $m \in \mathrm{S}^*_{\mathrm{RNS}}$ wird jeweils ein Löschexperiment durchgeführt. Dafür wird ein Differenzialgleichungssystem für $t = 200$ gelöst, das dem ungestörten entspricht, in dem $\frac{dx_m}{dt} \equiv 0$ gesetzt wurde. Der erhaltene Konzentrationsvektor wird mit $\mathbf{x}^{-\mathrm{m}}(t) =_{df} (x_1^{-m}(t), \ldots, x_n^{-m}(t))$ bezeichnet.

Die Durchführung dieser Simulationen erfolgte auf die in Abschnitt 4.3 besprochene Weise. Durch den erheblichen Rechenaufwand konnten nur jeweils zehn Genregulationssysteme untersucht werden. Die Lösung eines Differenzialgleichungssystems dieser Größe benötigte auf den verwendeten Rechnern zwischen einer halben und 4 Minuten, die Durchführung aller Löschexperimente eines Genregulationssystems mehrere Stunden.

Aus den so erhaltenen Konzentrationsvektoren des Wildtyps und der Störexperimente wurde die *durchschnittliche absolute Änderung der Konzentrationen* $\Delta_{\mathrm{abs}}\mathbf{x}(t)$ zum Zeitpunkt t wie folgt berechnet:

$$\Delta_{\mathrm{abs}}\mathbf{x}(t) \quad =_{df} \quad \frac{1}{l} \sum_{m \in \mathrm{S}^*_{\mathrm{RNS}}} \sum_{\substack{i=1 \\ i \neq m}}^{n} |x_i^{\mathrm{wt}}(t) - x_i^{-m}(t)| \quad . \tag{5.2}$$

Die jeweilige Änderung der Konzentration der gelöschten RNS wird hierbei nicht mitgemessen!

Abbildung 5.19 zeigt die durchschnittliche absolute Änderung der Konzentrationen für verschiedene mittlere multiple Eingangsgrade. Die absolute Änderung der Konzentrationen steigt mit wachsendem multiplen Eingangsgrad im untersuchten Intervall an. Mit größerem multiplem Eingangsgrad nimmt auch der mittlere Ausgangsgrad zu, sodass die Löschung eines Gens mehr Gene beeinflusst. Dem entgegen wirkt die zunehmende Anzahl redundanter Pfade und Rückkopplungen mit steigendem Verbindungsgrad. Es ist somit nicht zu erwarten, dass der in Abbildung 5.19 beobachtete Trend bei höheren multiplen Eingangsgraden fortgesetzt wird.

Wie vermutet, bewirkt die Löschung eines Gens eine geringe Änderung: Im System mit größtem $\Delta_{\mathrm{abs}}\mathbf{x}(200)$ beträgt die Länge des Konzentrationsvektors $\|\mathbf{x}^{\mathrm{wt}}(200)\|_1$ des Wildtyps etwa 190, wobei $\|(x_1(t), \ldots, x_n(t))\|_1 =_{df} \sum_{i=1}^{n} |x_i(t)|$. Dieser wird im Mittel über alle Löschexperimente um weniger als 3% ($\Delta_{\mathrm{abs}}\mathbf{x}(200) = 4.93$) verschoben.

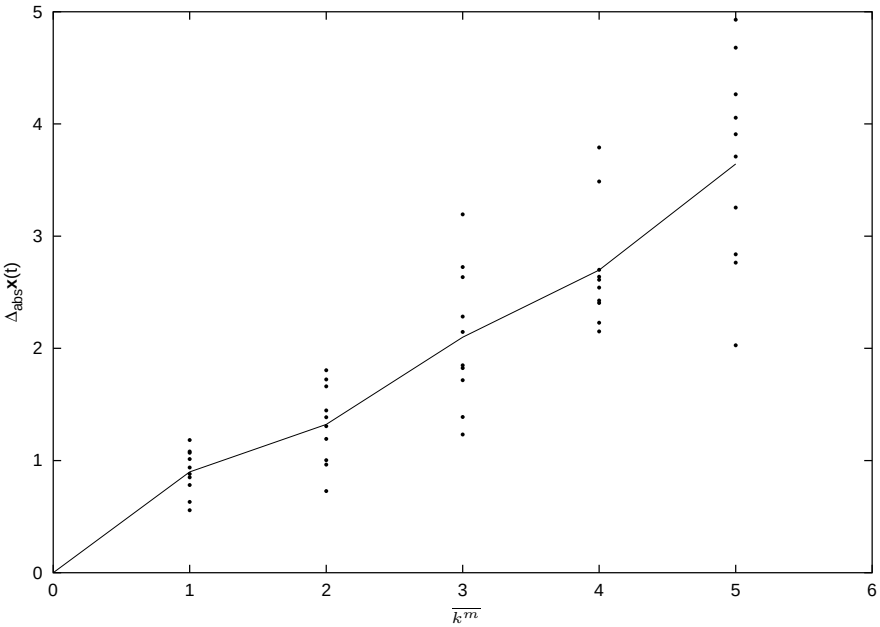

Abb. 5.19.: Durchschnittliche absolute Änderung der Konzentrationen $\Delta_{\text{abs}}\mathbf{x}(t)$ (Gleichung 5.2) zum Zeitpunkt $t = 200$ für jeweils zehn Genregulationssysteme in Abhängigkeit des mittleren multiplen Eingangsgrads $\overline{k^m}$. Die Linie verbindet die Mittelwerte. Dabei wird angenommen, dass bei einem mittleren Eingangsgrad von Null $\Delta_{\text{abs}}\mathbf{x}(t) = 0$ gilt. $N = 100$, $n_{\text{GRS}} = 1$, $k^m_{\max} = 2(\overline{k^m} + \sigma)$.

Neben der Gesamtänderung des Konzentrationsvektors kann auch die relative Änderung in jeder Dimension betrachtet werden: Die *durchschnittliche relative Änderung der Konzentrationen* $\Delta_{\text{rel}}\mathbf{x}(t)$ zum Zeitpunkt t sei definiert als:

$$\Delta_{\text{rel}}\mathbf{x}(t) \quad =_{df} \quad \frac{1}{(n-1)\times l} \sum_{m \in \mathrm{S}^*_{\text{RNS}}} \sum_{\substack{i=1 \\ i \neq m}}^{n} \frac{|x_i^{\text{wt}}(t) - x_i^{-m}(t)|}{|x_i^{\text{wt}}(t) + x_i^{-m}(t)|} \quad . \tag{5.3}$$

Bei diesem Maß wird zusätzlich auf die Anzahl der Dimensionen der Vektoren normiert.

In Abbildung 5.20 ist die durchschnittliche relative Änderung der Konzentrationen für verschiedene mittlere multiple Eingangsgrade dargestellt. Die durchschnittliche relative Änderung der Konzentrationen zeigt im untersuchten Intervall keine Abhängigkeit vom mittleren multiplen Eingangsgrad. Während die absolute Änderung eher ein Maß dafür ist, wieviele Substanzen durch eine Löschung betroffen sind, misst $\Delta_{\text{rel}}\mathbf{x}(t)$ die Stärke der Änderung der betroffenen Substanzen. Die Änderung im Konzentrationsverlauf einer Substanz durch Löschung einer einzelnen Kante im zugehörigen GRN-Graph hängt im verwendeten Modell nur vom Gewicht der Kante ab. Da die Gewichte des Differenzialgleichungssystems unabhängig vom multiplen Eingangsgrad gewählt werden, ist die relative Konzentrationsänderung bei Löschung einer einzelnen Kante unabhängig von der Verteilung des multiplen Eingangsgrads. Existieren multiple Kanten zwischen zwei Genen, so addieren sich die einzelnen Einflüsse. Mit zunehmendem Verbindungsgrad ist mit einer steigenden Anzahl multipler Kanten zwischen zwei Genen und somit mit einer höheren durchschnittlichen relativen Änderung der Konzentrationen zu rechnen. Diese Vorhersage konnte hier nicht beobachtet werden. Es ist einer Untersuchung wert, ob dieser Einfluss bei höheren mittleren multiplen Eingangsgraden zum Tragen kommt.

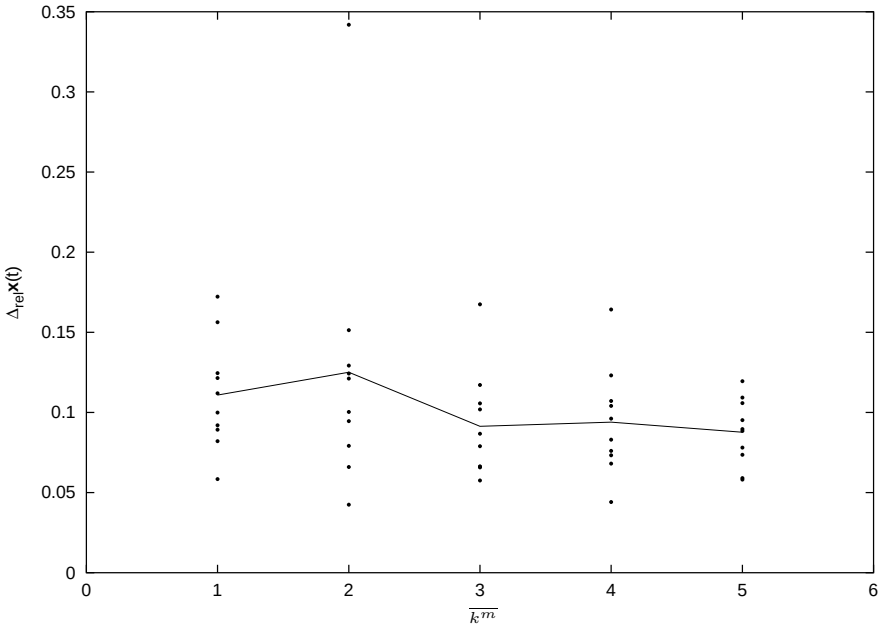

Abb. 5.20.: Durchschnittliche relative Änderung der Konzentrationen $\Delta_{\text{rel}}\mathbf{x}(t)$ (Gleichung 5.3) zum Zeitpunkt $t = 200$ für jeweils zehn Genregulationssysteme mit verschiedenem mittlerem multiplem Eingangsgrad $\overline{k^m}$. Die Linie verbindet die Mittelwerte. $N = 100$, $n_{\text{GRS}} = 1$, $k^m_{\text{max}} = 2(\overline{k^m}+\sigma)$.

Für ein Genregulationssystem mit $N = 100$ Knoten wurde aus allen einfachen Löschexperimenten dasjenige mit der größten absoluten Änderung der Konzentrationen bestimmt. In Abbildung 5.21 sind die Konzentrationsverläufe einiger ausgewählter Substanzen im Wildtyp und unter diesem Störexperiment dargestellt. Dabei wurden Substanzen ausgewählt, die eine große Änderung in den Konzentrationen aufweisen. Die Änderungen in den Konzentrationsverläufen der meisten anderen Substanzen sind sehr gering, wie das Histogramm der relativen Konzentrationsänderungen zeigt (Abbildung 5.22). Dabei ist die relative Konzentrationsänderung zwischen dem Wildtyp und der Löschung einer RNS $m \in \text{S}_{\text{RNS}}$ definiert als:

$$\Delta^m_{\text{rel}} x_s(t) \quad =_{df} \quad \frac{|x_s^{\text{wt}}(t) - x_s^{-m}(t)|}{|x_s^{\text{wt}}(t) + x_s^{-m}(t)|} \quad . \tag{5.4}$$

Für die gelöschte RNS m_{35} ist die relative Konzentrationsänderung per Definition gleich eins und für die in Abbildung 5.21 dargestellten Substanzen besitzt sie folgenden Werte:

$$\Delta^{m_{35}}_{\text{rel}} x_{m_{35}} = 1.00, \tag{5.5}$$
$$\Delta^{m_{35}}_{\text{rel}} x_{P_{25}} = 0.80, \tag{5.6}$$
$$\Delta^{m_{35}}_{\text{rel}} x_{P_{28}} = 0.26, \tag{5.7}$$
$$\Delta^{m_{35}}_{\text{rel}} x_{m_{49}} = 0.35, \tag{5.8}$$
$$\Delta^{m_{35}}_{\text{rel}} x_{m_{18}} = 0.26. \tag{5.9}$$

Über 70% der Substanzen haben eine relative Konzentrationsänderung zwischen dem Wildtyp und der Löschung von m_{35} von weniger als 5%.

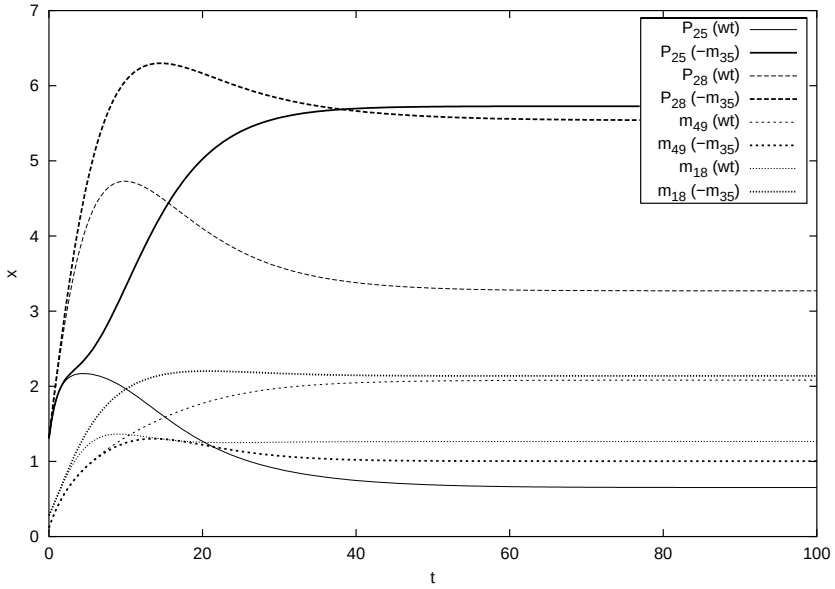

Abb. 5.21.: Konzentrationsverläufe ausgewählter Substanzen im Wildtyp (wt, dünne Linien) und nach der Löschung von m_{35} ($-m_{35}$, dicke Linien) in einem Genregulationssystem mit $N = 100$ Knoten ($n_{\mathrm{GRS}} = 1$). Jede Substanz ist in beiden Experimenten mit der gleichen Linienart, aber verschiedenen Dicken dargestellt.

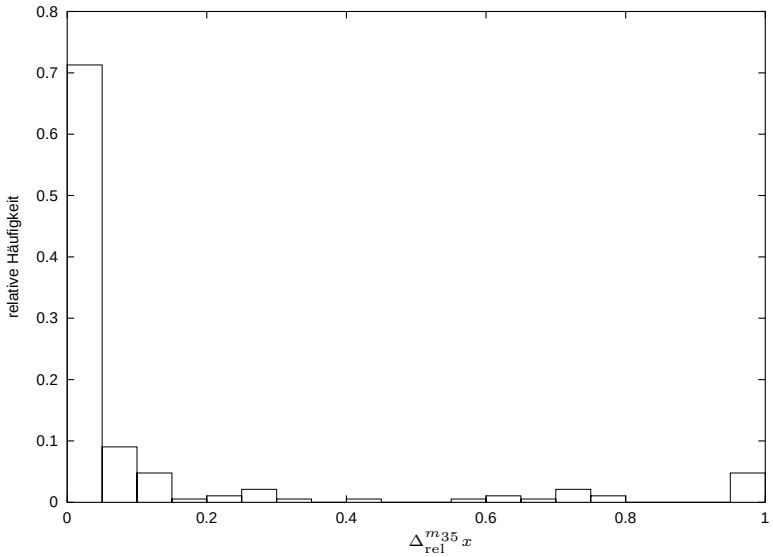

Abb. 5.22.: Histogramm der relativen Häufigkeit von Substanzen mit einer bestimmten relativen Konzentrationsänderung $\Delta_{\mathrm{rel}}^{m_{35}} x$ (Gleichung 5.4) zwischen Wildtyp und der Löschung von m_{35} zum Zeitpunkt $t = 300$ für das Genregulationssystem aus Abbildung 5.21.

6. Zusammenfassung

Mit dieser Arbeit ist ein flexibles Modell für die Genexpression entwickelt worden, das es ermöglicht, künstliche Genexpressionsdaten zu erzeugen. Diese Daten können zur Bewertung von Inferenzverfahren für Genregulationsnetze eingesetzt werden. Das Modell ist Grundlage für ein implementiertes Verfahren zur Generierung von Testdaten auf der Basis künstlicher Genregulationsnetze. Die Leistungsfähigkeit dieser Implementierung wird durch experimentelle Untersuchungen demonstriert.

Es hat sich gezeigt, dass die Bewertung von Inferenzverfahren häufig mit zu stark vereinfachten Modellen oder sehr kleinen Systemen erfolgt (Kapitel 3). Demgegenüber steht der in Kapitel 4 vorgestellte Modellrahmen, welcher sowohl Transkriptions- als auch die Translationvorgänge explizit berücksichtigt. Außerdem kann die Zusammenlagerung von Substanzen formuliert werden. Der Modellrahmen lässt die konkrete kinetische Beschreibung der Prozesse offen und ist damit flexibel genug, an Bedürfnisse hinsichtlich Rechenaufwand versus biologische Plausibilität angepasst zu werden.

Der vorgestellte Modellrahmen dient als Ausgangspunkt für ein Verfahren, mit dem künstliche Genregulationsnetze erzeugt werden können. Dabei sind Eigenschaften der künstlichen Netze durch Parameter vorgebbar. Im Kapitel 5 wird gezeigt, dass das entwickelte Programm in der Lage ist, künstliche Genregulationsnetze zu erzeugen, die den Vorgaben entsprechen. Diese künstlichen Genregulationsnetze werden benutzt, um Genexpressionsdaten zu generieren. Es können sowohl Zeitserien produziert, als auch Störexperimente durchgeführt werden. Das Verfahren gestattet die Erzeugung sehr großer künstlicher Genregulationsnetze mit vergleichsweise geringem Rechenaufwand (einige Sekunden für ein Netz mit 1000 Knoten). Die Generierung von Testdaten aus großen Genregulationsnetzen ist weniger effizient (einige Stunden für ein Netz mit 1000 Knoten).

Wie kann es weitergehen? Das implementierte Verfahren sollte zum einen in der Praxis eingesetzt werden, d.h. es sind Inferenzverfahren mit Hilfe des Testdatengenerators zu überprüfen. Zum anderen sind Modifikationen und Erweiterungen des Verfahrens denkbar: Künstliche Genregulationsnetze mit anderen strukturellen Eigenschaften, z.B. Verteilung des Eingangsgrads nach Potenzgesetz, könnten erzeugt werden. Die erwähnten alternativen Möglichkeiten zur Erzeugung künstlicher Genregulationsnetze könnten untersucht werden, so z.B. aus künstlichen Genomen (Reil, 1999) oder durch „Züchtung“ (Kauffman, 1993). Eine Implementierung anderer, vor allem nichtadditiver dynamischer Modelle, z.B. dem Modell von Yeung et al. (2002), könnte erfolgen. Um einen Einsatz des Verfahrens in der Praxis zu unterstützen, sollte eine zweckmäßige Benutzerschnittstelle geschaffen und die Verfügbarkeit im Internet ermöglicht werden. Außerdem sollte ein geeigneter Algorithmus zur Lösung der Differenzialgleichungen integriert werden.

Bei der Untersuchung der Inferenzverfahren für Genregulationsnetze und bei der Gestaltung des Modellrahmens für die Genexpression traten eine Reihe von Schwierigkeiten auf, die einer weiteren Untersuchung wert sind. Insbesondere fiel auf, dass zentrale Begriffe des Forschungsgebiets in verschiedener Bedeutung verwendet oder unscharf definiert werden. Interessante Fragen in diesem Zusammenhang sind:

1. Was repräsentiert eine Kante in einem Genregulationsnetz?
2. Welche Kanten stellen direkte, welche indirekte Einflüsse dar?
3. Was wird unter dem Eingangsgrad eines Knotens verstanden? Diese Frage hängt direkt mit der Definition eines Genregulationsnetzes zusammen.

Ob es möglich ist, den Traum von der Rekonstruktion von Genregulationsnetzen aus Genexpressionsdaten (Dutilh und Hogeweg, 1999) in die Wirklichkeit umzusetzen, ist eine offene Frage. Eine klare, einheitliche Beschreibung des zu lösenden Problems stellt einen notwendigen Schritt dar. Ebenso wichtig ist es, geeignete Testsysteme für die Inferenzverfahren zu schaffen. Die vorliegende Arbeit leistet einen Beitrag in beide Richtungen.

A. XML-Schemata

A.1. Schema für Parameterdateien

```
<?xml version="1.0"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
        xmlns:jxb="http://java.sun.com/xml/ns/jaxb">

<xsd:annotation>
<xsd:documentation xml:lang="en">
Title: Parameter schema for an artificial gene regulatory network
Author: Christian Knuepfer
Project: Generating artificial gene expression data
Organisation: CSB Jena, Germany
Date: 2003-01-08
</xsd:documentation>

<xsd:appinfo>
<jxb:globalBindings
        generateIsSetMethod="false"/>
</xsd:appinfo>
</xsd:annotation>

<xsd:element name="parameters" type="parametersType"/>

<xsd:element name="comment" type="xsd:string"/>

<xsd:complexType name="parametersType">
<xsd:all>
  <xsd:element name="description" minOccurs="0">
    <xsd:complexType>
       <xsd:sequence>
         <xsd:element name="date" type="xsd:date"/>
         <xsd:element name="time" type="xsd:time"/>
         <xsd:element name="ID" type="xsd:string"/>
       </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="numNodes" type="positiveInt"/>
  <xsd:element name="GeneProperties" type="GeneProps"/>
  <xsd:element name="fracSelfRegulators" type="D01" minOccurs="0" default="0.0"/>
  <xsd:element name="numHubs" type="positiveInt" minOccurs="0" default="0"/>
  <xsd:element name="fracExternRegulators" type="D01" minOccurs="0" default="0.0"/>
  <xsd:element name="substances" type="substancePortitions"/>
</xsd:all>
<xsd:attribute name="version" type="xsd:string"/>
</xsd:complexType>

<xsd:complexType name="substancePortitions">
<xsd:all>
  <xsd:element name="factorRNA" type="unsignedDouble"/>
  <xsd:element name="factorProtein" type="unsignedDouble"/>
  <xsd:element name="factorHeterodimer" type="unsignedDouble"/>
  <xsd:element name="factorHomodimer" type="unsignedDouble"/>
  <xsd:element name="dispDimerRNA" type="D01" minOccurs="0" default="1.0">
        <xsd:annotation>
        <xsd:appinfo>
                <jxb:property generateIsSetMethod="true"/>
        </xsd:appinfo>
        </xsd:annotation>
```

```
  </xsd:element>
  <xsd:element name="fracNoProduct" type="D01" minOccurs="0" default="0.0"/>
</xsd:all>
</xsd:complexType>

<xsd:complexType name="GeneProps">
<xsd:all>
  <xsd:element name="indegree" type="distributionType"/>
  <xsd:element name="dispRegulationRNA" type="D01" minOccurs="0" default="1.0">
  <xsd:annotation>
        <xsd:appinfo>
                <jxb:property generateIsSetMethod="true"/>
        </xsd:appinfo>
  </xsd:annotation>
  </xsd:element>
  <xsd:element name="fracTkrRegulation" type="D01" minOccurs="0" default="0.5">
  <xsd:annotation>
        <xsd:appinfo>
                <jxb:property generateIsSetMethod="true"/>
        </xsd:appinfo>
  </xsd:annotation>
  </xsd:element>
</xsd:all>
</xsd:complexType>

<xsd:complexType name="distributionType">
<xsd:choice>
  <xsd:element name="uniformDistribution" type="uniformDistributionType"/>
  <xsd:element name="normalDistribution" type="normalDistributionType"/>
</xsd:choice>
</xsd:complexType>

<xsd:complexType name="uniformDistributionType">
<xsd:all>
  <xsd:element name="minimum" type="unsignedInt" minOccurs="0" default="0"/>
  <xsd:element name="maximum" type="positiveInt"/>
</xsd:all>
</xsd:complexType>

<xsd:complexType name="normalDistributionType">
<xsd:all>
  <xsd:element name="maximum" type="positiveInt" minOccurs="0"/>
  <xsd:element name="mean" type="positiveDouble"/>
  <xsd:element name="sigma" type="unsignedDouble"/>
</xsd:all>
</xsd:complexType>

<!--simple numerical types-->
<xsd:simpleType name="positiveInt">
<xsd:restriction base="xsd:int">
<xsd:minExclusive value="0"/>
</xsd:restriction>
</xsd:simpleType>

<xsd:simpleType name="unsignedInt">
<xsd:restriction base="xsd:int">
<xsd:minInclusive value="0"/>
</xsd:restriction>
</xsd:simpleType>

<xsd:simpleType name="D01">
```

```
<xsd:restriction base="xsd:double">
<xsd:minInclusive value="0.0"/>
<xsd:maxInclusive value="1.0"/>
</xsd:restriction>
</xsd:simpleType>

<xsd:simpleType name="positiveDouble">
<xsd:restriction base="xsd:double">
<xsd:minExclusive value="0.0"/>
</xsd:restriction>
</xsd:simpleType>

<xsd:simpleType name="unsignedDouble">
<xsd:restriction base="xsd:double">
<xsd:minInclusive value="0.0"/>
</xsd:restriction>
</xsd:simpleType>

</xsd:schema>
```

A.2. Schema für GRN-Graphen

```
<?xml version="1.0"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">

<xsd:annotation>
<xsd:documentation xml:lang="en">
Title: Schema for the description of an artificial gene regulatory network
Author: Christian Knuepfer
Project: Generating artificial gene expression data
Organisation: CSB Jena, Germany
Date: 2003-01-08
</xsd:documentation>
</xsd:annotation>

<xsd:element name="GeneRegulatoryNetwork" type="GRNType"/>

<xsd:element name="comment" type="xsd:string"/>

<xsd:complexType name="GRNType">
<xsd:sequence>
        <xsd:element name="description">
                <xsd:complexType>
                <xsd:sequence>
                        <xsd:element name="date" type="xsd:date"/>
                        <xsd:element name="time" type="xsd:time"/>
                        <xsd:element name="ID" type="xsd:string"/>
                </xsd:sequence>
                </xsd:complexType>
        </xsd:element>
        <xsd:element name="Genom">
                <xsd:complexType>
                <xsd:sequence>
                        <xsd:element name="gene" type="geneType"
                         minOccurs="0" maxOccurs="unbounded"/>
                </xsd:sequence>
                </xsd:complexType>
        </xsd:element>
        <xsd:element name="GeneRegulatoryEdges">
                <xsd:complexType>
```

```
                <xsd:sequence>
                        <xsd:element name="edge" type="edgeType"
                         minOccurs="0" maxOccurs="unbounded"/>
                </xsd:sequence>
                </xsd:complexType>
        </xsd:element>
</xsd:sequence>
<xsd:attribute name="version" type="xsd:string"/>
</xsd:complexType>


<xsd:complexType name="geneType">
<xsd:sequence>
        <xsd:element name="name" type="xsd:string"/>
        <xsd:element name="RNA" type="xsd:string"/>
        <xsd:element name="product" type="xsd:string"/>
</xsd:sequence>
</xsd:complexType>

<xsd:complexType name="edgeType">
<xsd:sequence>
        <xsd:element name="label" type="xsd:string"/>
        <xsd:element name="from" type="xsd:string"/>
        <xsd:element name="to" type="xsd:string"/>
        <xsd:element name="strength" type="xsd:double"/>
</xsd:sequence>
</xsd:complexType>

</xsd:schema>
```

Literatur

Akutsu, T.; Kuhara, S.; Maruyama, O. und Miyano, S. (1998a). *Identification of Gene Regulatory Networks by Strategic Gene Disruptions and Gene Overexpressions*. In: *Proc. Ninth ACM-SIAM Symp. Discrete Algorithms (SODA '98)*, S. 695–702.

Akutsu, T.; Kuhara, S.; Maruyama, O. und Miyano, S. (1998b). *A System for Identifying Genetic Networks from Gene Expression Patterns Produced by Gene Disruptions and Overexpressions*. In: *Proc. Ninth ACM-SIAM Symp. Discrete Algorithms (SODA '98)*, Bd. 9, S. 151–160.

Akutsu, T.; Miyano, S. und Kuhara, S. (1999). *Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 4, S. 17–28.

Akutsu, T.; Miyano, S. und Kuhara, S. (2000). *Algorithms for Inferring Qualitative Models of Biological Networks*.

Alter, O.; Brown, P. O. und Botstein, D. (2000). *Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling*. In: *Proc. Natl. Acad. Sci. 97*, 18, S. 10101–10106.

Ando, S. und Iba, H. (2001). *Inference of Gene Regulatory Model by Genetic Algorithms*. In: *Proc. of the 2001 IEEE Congress on Evolutionary Computation*, S. 712–719. IEEE, Seoul, Korea.

Arnone, M. I. und Davidson, E. H. (1997). *The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems*. In: Development, 124: S. 1851–1864.

Barabási, A.-L. und Albert, R. (1999). *Emergence of Scaling in Random Networks*. In: Science, 286 (5439): S. 509–512.

Bernal, A.; Ear, U. und Kyrpides, N. (2001). *Genomes OnLine Database (GOLD): A Monitor of Genome Projects World-Wide*. In: Nucleic Acids Research, 29 (1): S. 126–127. URL http://ergo.integratedgenomics.com/GOLD/.

Bray, T.; Paoli, J.; Sperberg-McQueen, C. M. und Maler, E. (2000). *Extensible Markup Language (XML) 1.0 (Second Edition)*. W3C Recommendation 6 October 2000, World Wide Web Consortium (W3C). http://www.w3.org/TR/REC-xml.

Charniak, E. (1991). *Bayesian Networks without Tears*. In: AI Magazine, 12 (4): S. 50–63.

Chen, T.; Filkov, V. und Skiena, S. S. (1999a). *Identifying Gene Regulatory Networks from Experimental Data*. In: *ACM-SIGACT The Third Annual International Conference on Computational Moledular Biology (RECOMB99)*, S. 94–103.

Chen, T.; He, H. L. und Church, G. (1999b). *Modeling Gene Expression with Differential Equations*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 4, S. 29–40.

Davidson, E. H.; Rast, J. P.; Oliveri, P.; Ransick, A.; Calestani, C.; Yuh, C.-H.; Minokawa, T.; Amore, G.; Hinman, V.; Arenas-Mena, C.; Otim, O.; Brown, C. T.; Livi, C. B.; Lee, P. Y.; Revilla, R.; Rust, A. G.; jun Pan, Z.; Schilstra, M. J.; Clarke, P. J. C.; Arnone, M. I.; Rowen, L.; Cameron, R. A.; McClay, D. R.; Hood, L. und Bolouri, H. (2002). *A Genomic Regulatory Network for Development*. In: Science, 295: S. 1669–1678.

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

de Jong, H.; Geiselmann, J.; Hernandez, C. und Page, M. (2001). *Genetic Network Analyzer: A Tool for the Qualitative Simulation of Genetic Regulatory Networks*. Techn. Ber. RR 4262, INRIA.

D'haeseleer, P. (2000). *Reconstructing Gene Networks from Large Scale Gene Expression Data*. Dissertation, The University of New Mexico.

D'haeseleer, P. und Fuhrman, S. (1999). *Gene Network Inference Using a Linear, Additive Regulation Model*. Submitted to Bioinformatics.

D'haeseleer, P.; Liang, S. und Somogyi, R. (2000). *Genetic Network Inference: from Co-Expression Clustering to Reverse Engineering*. In: Bioinformatics, 16 (8): S. 707–726.

D'haeseleer, P.; Wen, X.; Fuhrman, S. und Somogyi, R. (1998). *Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data*. In: Information Processing in Cells and Tissues, S. 203–212.

D'haeseleer, P.; Wen, X.; Fuhrman, S. und Somogyi, R. (1999). *Linear Modeling of mRNA Expression Levels During CNS Development and Injury*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 4, S. 41–52.

Dutilh, B. E. und Hogeweg, P. (1999). *Gene Networks from Microarray Data: Analysis of Data from Microarray Experiments, the State of the Art in Gene Network Reconstruction*. Report Binf.1999.11.01, Bioinformatics, Utrecht University.

Eaton, J. W. (1998). *GNU Octave*. University of Wisconsin, Department of Chemical Engineering, Madison WI 53719. Software und Dokumentation sind im Internet erhältlich, http://www.octave.org/.

Eggenberger, P. (1997). *Evolving Morphologies of Simulated 3D Organisms Based on Differential Gene Expression*. In: *Proc. 4th European Conf. Artificial Life (ECAL97)*.

Friedman, N.; Linial, M.; Nachman, I. und Pe'er, D. (2000). *Using Bayesian Networks to Analyze Expression Data*. In: Journal Computational Biology, 7: S. 601–620.

Hagemann, R. (1991). *Allgemeine Genetik*. Gustav Fischer Verlag, Jena, 3. Aufl.

Hannon, G. J. (2002). *RNA interference*. In: Nature, 418: S. 244–251.

Hartemink, A. J.; Gifford, D. K.; Jaakkola, T. S. und Young, R. A. (2001). *Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 6, S. 422–433.

Hasty, J.; McMillen, D.; Isaacs, F. und Collins, J. J. (2001). *Computational Studies of Gene Regulatory Networks: in Numero Molecular Biology*. In: Nature Reviews, 2: S. 268–279.

Holter, N. S.; Maritan, A.; Cieplak, M.; Fedoroff, N. V. und Banavar, J. R. (2001). *Dynamic Modeling of Gene Expression Data*. In: *Proc. Natl. Acad. Sci. 98*, 4, S. 1693–1698.

Ideker, T. E.; Thorsson, V. und Karp, R. M. (2000). *Discovery of Regulatory Interactions through Perturbation: Inference and Experimental Design*. In: *Proc. Pacific Symp. Biocomputing*, S. 305–316.

Imoto, S.; Goto, T. und Miyano, S. (2002). *Estimation of Genetic Networks and Functional Structures Between Genes by Using Bayesian Networks and Nonparametric Regression*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 7, S. 175–186.

Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. und Barabási, A.-L. (2000). *The Large-Scale Organization of Metabolic Networks*. In: Nature, 407: S. 651–654.

Kauffman, S. A. (1969). *Metabolic Stability and Epigenesis in Randomly Connected Nets*. In: Journal of Theoretical Biology, 22: S. 437–467.

Kauffman, S. A. (1993). *The Origins of Order – Self-Organization and Selection in Evolution*. Oxford University Press, Inc., New York, Oxford.

Kitano, H. (Hg.) (2000). *Foundations of Systems Biology*. MIT Press, Cambridge, MA.

Kyoda, K. M.; Morohashi, M.; Onami, S. und Kitano, H. (2000). *A Gene Network Inference Method from Continuous-Value Gene Expression Data of Wild-Type and Mutants*. In: Genome Informatics, 11: S. 196–204.

Lander, E. S. (1996). *The New Genomics: Global Views of Biology*. In: Science, 274 (5287): S. 536–539.

Lehninger, A. L.; Nelson, D. L. und Cox, M. M. (1994). *Prinzipien der Biochemie*. Spektrum Akad. Verl., Heidelberg, 2. Aufl. Deutsche Übersetzung herausgegeben von Harald Tschesche.

Liang, S.; Fuhrman, S. und Somogyi, R. (1998). *REVEAL, a General Reverse Engineering Algorithm for Inference of Genetic Network Architectures*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 3, S. 18–29.

Maki, Y.; Tominaga, D.; Okamoto, M.; Watanabe, S. und Eguchi, Y. (2001). *Development of a System for the Inference of Large Scale Genetic Networks*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 6, S. 446–458.

Mendes, P. (2000). *Modeling Large Biological Systems from Functional Genomic Data: Parameter Estimation*. In: Kitano (2000), Kap. 8, S. 163–186.

Mestl, T.; Plahte, E. und Omholt, S. W. (1995). *A Mathematical Framework for Describing and Analysing Gene Regulatory Networks*. In: Journal of Theoretical Biology, 176: S. 291–300.

Mjolsness, E.; Mann, T.; Castano, R. und Wold, B. (1999). *From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regulation among Gene Classes from Large-Scale Expression Data*. Techn. Ber. JPL-ICTR-99-4, Jet Propulsion Laborator, NASA.

Morohashi, M. und Kitano, H. (1999). *Identifying Gene Regulatory Networks from Time Series Expression Data by* in silico *Sampling and Screening*. In: *Proc. 5th European Conference on Artificial Life (ECAL '99)*, S. 477–486.

Munk, K. (Hg.) (2001). *Genetik*. Grundstudium Biologie. Spektrum Akademischer Verlag GmbH, Heidelberg, Berlin. Autoren: Matthias Fladung, Nina Krauß, Inge Kronberg, Thomas Langer, Gerlinde Linne von Berg, Katharina Munk, Regina Nethe-Jaenchen, Gunvor Pohl-Apel, Harald Schlatter und Klaus W. Wolf.

Onami, S.; Kyoda, K. M.; Morohashi, M. und Kitano, H. (2000). *The DBRF Method for Inferring a Gene Network from Large-Scale Steady-State Gene Expression Data*. In: Kitano (2000), Kap. 3, S. 59–75.

Pe'er, D.; Regev, A.; Elidan, G. und Friedman, N. (2001). *Inferring Subnetworks from Perturbed Expression Profiles*. In: Bioinformatics, 1 (1): S. 1–9.

Raychaudhuri, S.; Stuart, J. M. und Altman, R. B. (2000). *Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 5, S. 452–463.

Reil, T. (1999). *Dynamics of Gene Expression in an Artificial Genome - Implications for Biological and Artificial Ontogeny*. http://users.ox.ac.uk/~quee0818/agenome.ps.

Reil, T. (2000). *Models of Gene Regulation - A Review*. http://users.ox.ac.uk/~quee0818/genreg.ps.

Sander, G. (1995). *VCG – Visualization of Compiler Graphs*. Technical Report A01-95, Universität des Saarlandes, FB 14 Informatik. Software und Dokumentation sind im Internet erhältlich, ftp://ftp.cs.uni-sb.de/pub/graphics/vcg.

Sauermost, R. (Hg.) (1991-1992). *Lexikon der Biochemie und Molekularbiologie*. Verlag Herder, Freiburg im Preisgau.

Sauermost, R. (Hg.) (1994). *Herder-Lexikon der Biologie*. Spektrum Akademischer Verlag GmbH, Heidelberg, Berlin, Oxford. 8 Bände, Indexband und Ergänzungsband.

Savageau, M. (1998). *Rules for the Evolution of Gene Circuitry*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 3, S. 54–65.

Smolen, P.; Baxter, D. A. und Byrne, J. H. (2000). *Modeling Transcriptional Control in Gene Networks – Methods, Recent Results, and Future Directions*. In: Bulletin of Mathematical Biology, 62 (2): S. 247–292.

Somogyi, R. und Sniegoski, C. A. (1996). *Modeling the Complexity of Genetic Networks: Understanding Multigenic and Pleiotropic Regulation*. In: Complexity, 1 (6): S. 45–63.

Stadler, P. F.; Fontana, W. und Miller, J. H. (1993). *Random Catalytic Reaction Networks*. In: Physica D, 63: S. 378–392.

Szallasi, Z. (1999). *Genetic Network Analysis in Light of Massively Parallel Biological Data Acquisition*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 4, S. 5–16.

Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J. und Church, G. M. (1999). *Systematic Determination of Genetic Network Architecture*. In: Nature Genetics, 22: S. 281–285.

Thieffry, D.; Huerta, A. M.; Pérez-Rueda, E. und Collado-Vides, J. (1998). *From Specific Gene Regulation to Genomic Networks: a Global Analysis of Transcriptional Regulation in Escherichia Coli*. In: BioEssays, 20: S. 433–440.

Thieffry, D. und Thomas, R. (1998). *Qualitative Analysis of Gene Networks*. In: *Proc. Pacific Symp. Biocomputing*, S. 77–88.

Toh, H. und Horimoto, K. (2002). *Inference of a Genetic Network by a Combined Approach of Cluster Analysis and Graphicla Gaussian Modeling*. In: Bioinformatics, 18 (2): S. 287–297.

Voit, E. O. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge University Press, Cambridge.

Wahde, M. und Hertz, J. (2000). *Coarse-Grained Reverse Engineering of Genetic Regulatory Networks*. In: Biosystems, 55: S. 129–136.

Weaver, D.; Workman, C. und Stormo, G. (1999). *Modeling Regulatory Networks with Weight Matrices*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 4, S. 112–123.

Wessels, L.; Someren, E. V. und Reinders, M. (2001). *A Comparison of Genetic Network Models*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 6, S. 508–519.

Yeung, M. K. S.; Tegnér, J. und Collins, J. J. (2002). *Reverse Engineering Gene Networks Using Singular Value Decomposition and Robust Regression*. In: *Proc. Natl. Acad. Sci. 99*, S. 6163–6168.

Yoo, C.; Thorsson, V. und Cooper, G. (2002). *Discovery of Causal Relationships in a Gene-Regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data*. In: *Proc. Pacific Symp. Biocomputing*, Bd. 7, S. 498–509.

Yuh, C.-H.; Bolouri, H. und Davidson, E. H. (1998). *Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene*. In: Science, 279: S. 1896–1902.

Erklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Jena, den 22. März 2003 .....
(Christian Knüpfer)