Institut für Informatik
der Naturwissenschaftlichen Fakultät III
der
Martin-Luther-Universität Halle-Wittenberg

Zentralitätsanalyse molekularbiologischer Netzwerke

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt von

Diplom-Informatiker Dirk Koschützki
geb. am 29.07.1971 in Hamburg

Gutachter:

Prof. Dr. habil. Falk Schreiber

Prof. Dr. habil. Ralf Hofestädt

Verteidigung am:

14. 7. 2011

Furtwangen, März 2011

Inhaltsverzeichnis

Danksagung

Es gibt zwei Personen, die eine mehr als bedeutende Rolle während der gesamten Zeit, die ich an dieser Dissertation verbracht habe, gespielt haben: Meine Ehefrau Dipl.-Ing. Petra Wardzichowski und mein Betreuer Prof. Dr. Falk Schreiber. Beiden gebührt an dieser Stelle ein mehr als herzliches Dankeschön!

Eine sehr große Zahl an Freunden, Verwandten, Kolleginnen und Kollegen standen zu vielen Zeitpunkten mit Rat und Tat zur Seite. Auch ihnen ein herzliches Dankeschön.

Meine drei Kinder haben sicherlich *nicht* dazu beigetragen, dass diese Dissertation „fristgerecht“ fertig geworden ist. Dennoch möchte ich allen dreien an dieser Stelle ebenfalls herzlich danken!

Dirk Koschützki

1. Einleitung

> One of the primary uses of graph theory in social network analysis is the identification of the most important actors in a social network.[168]

Das vorstehende Zitat aus einem Standardwerk über die Analyse sozialer Netzwerk fasst den Nutzen der Zentralitätsanalyse für soziale Netzwerke sehr anschaulich zusammen: Durch die Analyse sollen u.a. die wichtigsten Beteiligten in einem Beziehungsnetzwerk identifiziert werden.

Die zugrunde liegende Methode für die Unterscheidung der wichtigen von den unwichtigen Individuen ist die *Zentralitätsanalyse*. Bei dieser Analyse wird allein auf der Basis der Struktur des zu untersuchenden sozialen Netzwerkes jedem einzelnen Individuum ein Wert, der die Wichtigkeit desselben beschreiben soll, zugeordnet. Beispielsweise kann davon ausgegangen werden, dass Individuen, die häufig mit anderen Individuen kommunizieren wichtiger sind, als Individuen, die nur wenige Kommunikationsverbindungen eingehen. Durch das Zählen der Kommunikationsverbindungen lässt sich folglich eine Reihenfolge der Individuen erstellen.

Die Methode an sich, d.h. die Festlegung einer Reihenfolge der Wichtigkeit von Objekten auf der Basis der Struktur eines zu untersuchenden Netzwerkes, wird jedoch auch in anderen Bereichen angewendet. Prominentestes Beispiel ist die Suchmaschine *Google*: Die Reihenfolge, in der die Resultate einer Suche dem Benutzer präsentiert werden, basiert auf der Struktur des durch die Verlinkung der Webseiten entstehenden Netzwerkes.

Molekularbiologische Netzwerke, d.h. die Repräsentation von molekularbiologischen Objekten und ihrer Verbindungen durch Knoten und Kanten in Form eines Graphen, sind seit vielen Jahren aktiver Gegenstand der Forschung. Grundlage hierfür bildet die Erkenntnis, dass die Betrachtung eines einzelnen molekularbiologischen Objekts, beispielsweise eines Proteins, als für sich alleine betrachtete Einheit, nur begrenzte Informationen über die Wirkungsweise desselben liefert. Ein Transkriptionsfaktor, d.h. ein Protein, dass die „Erstellung“ anderer Proteine reguliert bzw. beeinflusst, kann in der Regel nur in Kombination mit anderen Proteinen seine Wirkung entfalten. Die Funktionsweise des untersuchten Proteins kann also nur durch die Betrachtung der Wechselwirkungen mit anderen Proteinen verstanden werden.

Zu den molekularbiologischen Netzwerken gehören Genregulationsnetze, metabolische Reaktionsnetze, Protein-Interaktionsnetze und Signal-Transduktionsnetze. Genregulationsnetze veranschaulichen die gegenseitige Aktivierung bzw. Deaktivierung einzelner Gene untereinander. In metabolischen Reaktionsnetzen werden die Abläufe biochemischer Reaktionen modelliert. Protein-Interaktionsnetze beschreiben die Wechselwirkungen zwischen einzelnen Proteinen. Die Erkennung und Übermittlung von Reizen, beispielsweise dem Vorhandensein eines Hormons, wird in Form von Signal-Transduktionsnetzen dargestellt.

Im Kontext der Analyse molekularbiologischer Netzwerke lassen sich die durch eine Zentralitätsanalyse ermittelte Bewertung der Knoten für die Beantwortung einer ganzen Reihe von Fragestellungen nutzen:

Bestimmung von Netzwerkelementen mit spezifischen Eigenschaften Mit Hilfe einer Bewertung der Knoten lassen sich Knoten mit gewünschten Eigenschaften ermitteln. Durch eine Zentralitätsanalyse können beispielsweise Knoten, die potentiell eine weitreichende Kontrolle ausüben können, identifiziert werden.

Steuerung von Experimenten Die Analyse einzelner molekularbiologischer Objekte im Rahmen von Experimenten ist zeitaufwändig und kostenintensiv. Eine mögliche Vorauswahl der zu untersuchenden Objekte auf der Basis einer vorher erstellten Reihenfolge hilft somit Kosten zu sparen und potentiell zeitraubende Untersuchungen von „uninteressanten“ Objekten zu vermeiden.

Exploration von Netzwerken Die in einem molekularbiologischen Netzwerk beschriebenen Informationen sind ausgesprochen umfangreich. Eine Einarbeitung in ein unbekanntes molekularbiologisches Netzwerk ist folglich ein zeitraubender Prozess. Eine Bewertung und Sortierung der einzelnen Objekte kann diesen Prozess vereinfachen und beschleunigen.

Zurzeit sind über 20 verschiedene *Zentralitätsmaße* (auch *Zentralitäten*) zur Analyse von Netzwerken bekannt. Jedes dieser Zentralitätsmaße verwendet bei der Bestimmung der Wichtigkeit der Knoten einen festgelegten Bewertungsmaßstab. Dieser Bewertungsmaßstab besagt beispielsweise, dass ein Knoten, der viele Verbindungen zu anderen Knoten hat, ein wichtiger Knoten innerhalb des modellierten Systems ist. Die Frage, welches Zentralitätsmaß für die Analyse eines bestimmten Typs von Netzwerken einzusetzen ist, ist dabei einerseits vom Netzwerktyp und andererseits von der zu beantwortenden Fragestellung abhängig.

Für die Analyse von molekularbiologischen Netzwerken ergibt sich das Problem, dass die ungeprüfte Anwendung existierender Zentralitäten möglicherweise nicht aussagekräftige Resultate liefert. Bereits an einem einfachen Beispiel wird dieses deutlich: Die Zentralität *Closeness* wurde bereits mehrfach zur Analyse von molekularbiologischen Netzwerken eingesetzt. Eine der dieser Zentralität zugrunde liegenden Annahmen ist, dass der durch das Netzwerk beschriebene Prozess ausschließlich über kürzeste Wege abläuft. Dass diese Annahme für Genregulationsnetze zutrifft, lässt sich noch einsehen: Der in diesen Netzen modellierte Prozess beschreibt die Aktivierung eines Gens durch den zugehörigen Transkriptionsfaktor. Die Annahme, dass eine Aktivierung eines Gens durch ein anderes immer auf kürzestem Wege erfolgt ist zumindest plausibel. Für Protein-Interaktionsnetze ist diese Annahme allerdings bereits fragwürdig: Eine Interaktion zwischen zwei Proteinen kann nur entweder erfolgen oder nicht. Die Wichtigkeit eines Proteins hängt deshalb möglicherweise von der Anzahl der Interaktionspartner ab. Warum hingegen die (graphentheoretische) Länge der Wege zwischen interagierenden Proteinen einen Einfluss auf die Wichtigkeit von Proteinen haben soll, ist momentan nicht bekannt. Spätestens aber bei der Betrachtung von metabolischen Reaktionsnetzen wird offensichtlich, dass diese Betrachtungsweise irreführend ist. Gerade das Vorhandensein von Reaktionswegen, die *nicht* optimal im Sinne des kürzesten Weges sind, führt zur Flexibilität und Robustheit metabolischer Reaktionsnetze und somit zur Überlebensfähigkeit bzw. zur Möglichkeit der flexiblen Anpassung eines Organismus an Umwelteinflüsse. Bereits aus dieser Beobachtung leitet sich die Forderung nach der Entwicklung von speziellen Zentralitätsmaßen für die Analyse molekularbiologischer Netzwerke ab. Diese Zentralitätsmaße sollten dabei die in einer Zelle ablaufenden molekularbiologischen Prozesse bei der Festlegung einer Reihenfolge der Objekte angemessen berücksichtigen.

In dieser Dissertation werden deshalb zwei neue Zentralitäten für molekularbiologische Netzwerke entwickelt. Die erste, *Motiv-basierte Zentralität* genannt, ist besonders für die

Analyse von Genregulationsnetzen geeignet und die zweite, *Fluss-basierte Zentralität* genannt, für die Analyse von metabolischen Reaktionsnetzen. Die *Motiv-basierte Zentralität* nutzt zur Bestimmung der Zentralitätswerte das Vorkommen von Motiven, d.h. kleinen, häufig im Netzwerk auftauchenden Teilnetzwerken, aus. Diese Motive sind in Genregulationsnetzen bereits als wichtige Bausteine identifiziert worden. Insbesondere globale Regulatoren, d.h. Gene, die einen großen regulatorischen Einfluss auf andere Gene haben, lassen sich mit Hilfe dieser neuen Zentralität sehr gut identifizieren. Die *Fluss-basierte Zentralität* dient zur Analyse metabolischer Flussnetze. Metabolische Flussnetze sind metabolische Reaktionsnetze, in denen die Kanten mit dem metabolischen Fluss, d.h. der Menge der transportierten bzw. umgesetzten Stoffmenge, gewichtet sind. Mittels dieser Zentralität lassen sich Metaboliten, die in viele andere Metaboliten umgewandelt werden können, identifizieren.

Diese Dissertation ist wie folgt strukturiert: In Kapitel 2 werden die mathematischen Grundlagen, d.h. die notwendigen Begriffe aus der Graphentheorie und die verwendeten Zentralitäten, eingeführt. Kapitel 3 umfasst die für das Verständnis der Dissertation notwendigen molekularbiologischen Grundlagen. In Kapitel 4 wird, basierend auf den bereits publizierten Zentralitätsanalysen, die Notwendigkeit für neue Zentralitätsmaße motiviert. Die Motiv-basiert Zentralität für die Analyse von Genregulationsnetzen wird im Kapitel 5 vorgestellt und die Bewertungen der einzelnen Gene in Hinblick auf die Identifikation von globalen Regulatoren werden diskutiert. Für die Analyse von metabolischen Reaktionsnetzen wird im Kapitel 6 die Fluss-basierte Zentralität definiert und Reihenfolgen von Metaboliten auf der Basis dieser Zentralität werden betrachtet. Kapitel 7 schließt diese Dissertation mit einer Zusammenfassung ab.

2. Graphen und Zentralitätsmaße

Molekularbiologische Netzwerke werden als Graphen modelliert und können mit Hilfe von Zentralitätsmaßen analysiert werden. Im Folgenden werden deshalb alle notwendigen Begriffe aus der Graphentheorie definiert und die zur Analyse von molekularbiologischen Netzwerken geeigneten bzw. bereits eingesetzten Zentralitätsmaße vorgestellt.

2.1. Graphen

Die Definitionen folgen im Wesentlich der Notation aus Kapitel 2 eines Netzwerkanalyse-Buchs herausgegeben von U. Brandes & T. Erlebach [24] und dem Graphentheorie-Buch von R. Diestel [38].

Definition 2.1 (Gerichteter Graph, Gerichtete Kante, Start-/Endknoten)
Ein gerichteter Graph *ist ein Tupel* $G = (V, E)$ *bestehend aus einer endlichen Menge* $V \neq \emptyset$ *von* Knoten *und einer endlichen Menge* $E \subseteq V \times V$ *von* Kanten. *Jede Kante* $(e = (v_1, v_2))$ *ist* gerichtet *und die beiden Komponenten werden* Startknoten (v_1) *und* Endknoten (v_2) *genannt.*

Zu einem gegebenen Graphen $G = (V, E)$ bezeichnet $V(G)$ die *Knotenmenge* und $E(G)$ die *Kantenmenge*. Gemäß obiger Definition sind *Mehrfachkanten*, d.h. Kanten zwischen zwei Knoten mit derselben Richtung, nicht erlaubt. *Schleifen*, d.h. Kanten von einem Knoten zum selben Knoten, sind gemäß obiger Definition erlaubt.

Ein *ungerichteter Graph* ist ein Graph bei dem die Richtung der Kanten nicht unterschieden wird. Ungerichtete Graphen werden im Rahmen dieser Dissertation nur an wenigen Stellen verwendet und an all diesen können sie durch einen zugehörigen gerichteten Graphen repräsentiert werden. In diesem zugehörigen gerichteten Graphen wird jede ungerichtete Kante aus dem ungerichteten Graphen durch zwei antiparallele (gerichtete) Kanten modelliert. Schleifen im ungerichteten Graphen werden dabei durch eine Kante im gerichteten Graphen modelliert. Die im Folgenden für ungerichtete Graphen verwendeten Begriffe und Eigenschaften können durch diese Modellierung auf die für gerichtete Graphen definierten Begriffe zurückgeführt werden.

Im Folgenden wird die Bezeichnung *Graph* synonym für gerichteter Graph verwendet. Nur wenn eine Eigenschaft explizit für ungerichteten Graphen verwendet wird, dann wird diese Unterscheidung vorgenommen.

Definition 2.2 (Gewichteter Graph, Gewichtsfunktion)
Sei $G = (V, E)$ *ein Graph. Eine Funktion* $\omega \colon E \mapsto \mathbb{R}$ *wird* Gewichtsfunktion *zum Graphen* G *genannt. Der Graph* G *zusammen mit einer Gewichtsfunktion* ω *wird* gewichteter Graph *(geschrieben* (G, ω)*) genannt.*

Ein *ungewichteter Graph*, d.h. ein Graph ohne Gewichtsfunktion, kann als gewichteter Graph mit der trivialen Gewichtsfunktion $\forall e \in E \colon \omega(e) = 1$ betrachtet werden. Im Kontext

der Zentralitätsanalyse sind die Graphen üblicherweise entweder ungewichtet, wie beispielsweise die Genregulationsnetze (siehe Kapitel 5) oder die Kantengewichte sind strikt positiv, wie beispielsweise bei den Flussnetzen (siehe Kapitel 6).

Je nach Art des modellierten Prozesses bzw. betrachteten Systems können Kantengewichte eine unterschiedliche Bedeutung haben. Bei der Berechnung von kürzesten Wegen innerhalb eines Graphen führt beispielsweise ein geringes Kantengewicht dazu, dass die betreffende Kante häufig verwendet wird. Für den Fall, dass ein Fluss durch einen Graphen ermittelt wird, führt hingegen ein hohes Kantengewicht zu einem hohen Fluss über die betreffende Kante. Je nach Interpretation der Kantengewichte ist somit ein hohes oder niedriges Gewicht von Vorteil.

Definition 2.3 (Gelabelter Graph, Labelfunktion)
Sei A eine endliche Menge von Bezeichnern und sei $G = (V, E)$ ein Graph. Eine Funktion $l\colon V \mapsto A$ wird Labelfunktion *genannt. Ein Graph zusammen mit einer Labelfunktion wird* gelabelter Graph *genannt.*

Eine Labelfunktion gestattet die Benennung der Knoten.

Definition 2.4 (Teilgraph, induzierter Teilgraph)
Ein Graph $G' = (V', E')$ ist ein Teilgraph *eines anderen Graphen $G = (V, E)$ (geschrieben $G' \subseteq G$) genau dann, wenn $V' \subseteq V$ und $E' \subseteq E \cap (V' \times V')$. Ein Teilgraph $G' = (V', E')$ zu einem Graphen $G = (V, E)$ heißt* induzierter Teilgraph *genau dann, wenn für die Menge der Kanten $E' = E \cap (V' \times V')$ gilt.*

Definition 2.5 (Isomorpher Graph, Graph-Isomorphismus, -Automorphismus)
Seien $G_A = (V_A, E_A)$ und $G_B = (V_B, E_B)$ Graphen. Der Graph G_A heißt isomorph *zu G_B (geschrieben $G_A \simeq G_B$), genau dann, wenn eine Bijektion $\phi\colon V_A \mapsto V_B$ existiert für die gilt $\forall v_1, v_2 \in V_A\colon (v_1, v_2) \in E_A \Leftrightarrow (\phi(v_1), \phi(v_2)) \in E_B$. Die Abbildung ϕ wird* Graph-Isomorphismus *(oder nur* Isomorphismus*) genannt. Falls $G_A = G_B$ gilt, dann wird die Abbildung* (Graph-)Automorphismus *genannt.*

Ein Automorphismus für einen Graphen G beschreibt die „Selbstähnlichkeit“ des Graphen und die Automorphismen des Graphen bilden eine Gruppe, die *Automorphismengruppe* ($Aug(G)$) von G. Knoten, die durch einen Automorphismus aufeinander abgebildet werden, können anhand der Graphstruktur nicht voneinander unterschieden werden und bilden einen gemeinsamen *Orbit* ($Orb(v_1) := \{v_2 \in V_G \mid v_1 = g(v_2), g \in Aut(G)\}$).

Definition 2.6 (Benachbart, Vorgänger, Nachfolger)
Sei $G = (V, E)$ ein Graph. Zwei Knoten $v_1, v_2 \in V$ sind benachbart *genau dann, wenn eine Kante $e \in E$ mit $e = (v_1, v_2)$ existiert. Die Knoten, die mit einem Knoten v_1 durch eingehende bzw. ausgehende Kanten verbunden sind, werden* Vorgänger *bzw.* Nachfolger *von v_1 genannt. Die dazugehörige Menge ist definiert als $\Gamma^-(v_1) := \{v_0 \mid (v_0, v_1) \in E\}$ bzw. $\Gamma^+(v_1) := \{v_2 \mid (v_1, v_2) \in E\}$.*

Definition 2.7 (Weg, Pfad, Geschlossener Weg)
Sei $G = (V, E)$ ein Graph und $(e_1, \ldots, e_n)$ eine Folge von Kanten aus E. Wenn zu dieser Folge eine Folge von Knoten $(v_0, \ldots, v_n)$ aus V mit der Eigenschaft $e_i = (v_{i-1}, v_i), i \in \{1, \ldots, n\}$ existiert, dann heißt diese Folge von Kanten Weg *im Graph G. Wenn alle Knoten v_0 bis v_n paarweise verschieden sind, dann wird dieser Weg* Pfad *genannt. Ein* geschlossener Weg *ist ein Weg für den $v_0 = v_n$ gilt.*

Definition 2.8 (Länge eines Weges, Kürzester Pfad, Distanz)
Sei $G = (V, E)$ ein Graph und $\omega\colon E \mapsto \mathbb{R}$ eine Gewichtsfunktion zu G. Die Summe der Kantengewichte in einem Weg $w = (e_1, \ldots, e_n)$ wird Länge des Weges *($l(w) := \sum_{i=1}^{n} \omega(e_i)$) genannt. Ein Pfad zwischen zwei Knoten in einem Graphen G ist ein* kürzester Pfad *genau dann, wenn die beiden Knoten nicht durch einen Pfad mit geringerer Länge verbunden werden können. Die Länge eines kürzesten Pfades zwischen zwei Knoten v_1 und v_2 in einem Graph G heißt* Distanz *($\mathrm{dist}(v_1, v_2)$). Falls zwischen zwei Knoten v_1 und v_2 kein Weg existiert, dann gilt $\mathrm{dist}(v_1, v_2) := \infty$.*

Definition 2.9 (Zusammenhängend)
Ein ungerichteter Graph bei dem je zwei Knoten durch einen Weg verbunden sind heißt zusammenhängend.

Definition 2.10 (Stark zusammenhängend)
Ein gerichteter Graph $G = (V, E)$ heißt stark zusammenhängend *genau dann, wenn zwischen je zwei Knoten aus V mindestens ein Weg existiert.*

Bei der Eigenschaft stark zusammenhängend ist zu beachten, dass in einem gerichteten Graphen die Richtung der Kanten berücksichtigt werden muss.

Definition 2.11 (Zugehöriger ungerichteter Graph)
Der ungerichtete Graph $G_u = (V, E_u)$ der aus einem gerichteten Graphen $G = (V, E)$ entsteht, wenn zu jeder Kante aus E die entsprechende antiparallele Kante hinzugefügt wird ($E_u = E \cup \{(v_2, v_1) \mid (v_1, v_2) \in E\}$), heißt zugehöriger ungerichteter Graph.

Definition 2.12 (Schwach zusammenhängend)
Wenn der gerichtete Graph G nicht stark zusammenhängend ist, der zum Graph G zugehörige ungerichtete Graph G_u hingegen zusammenhängend ist, dann ist der gerichtete Graph G schwach zusammenhängend.

Definition 2.13 (Zusammenhangskomponente)
Ein Teilgraph G' zu einem gegebenen Graphen $G = (V, E)$ heißt starke (bzw. schwache) Zusammenhangskomponente, *falls G' stark (bzw. schwach) zusammenhängend und maximal ist.*

Maximal bedeutet, dass es keinen Teilgraphen G'' von G gibt, der stark (bzw. schwach) zusammenhängend ist und für den gilt $V'' \supset V'$.

Definition 2.14 (Adjazenzmatrix)
Ein Graph $G = (V, E)$ mit einer n-elementigen Knotenmenge ($|V| = n$) kann durch eine $n \times n$-Matrix ($A = (a_{ij})$) repräsentiert werden. In dieser Matrix ist der Eintrag an der Position (i, j) genau dann eins, wenn eine Kante von i nach j existiert, sonst ist dieser Eintrag 0. Diese Matrix A wird Adjazenzmatrix *von G genannt.*

Für gewichtete Graphen mit Gewichtsfunktion ω wird als Eintrag an der Position i, j das Kantengewicht $\omega(e)$ der Kante $e = (v_i, v_j)$ vermerkt.

Definition 2.15 (Inverser Graph)
Sei $G = (V, E)$ ein Graph. Der Graph $G_{inv} = (V_{inv}, E_{inv})$ mit $V_{inv} = V$ und $E_{inv} = \{(v_2, v_1) \mid (v_1, v_2) \in E\}$ wird inverser Graph *zum Graph G genannt.*

Der inverse Graph zu einem Graphen G entsteht also durch das „Umdrehen“ der Kantenrichtungen aller Kanten von G.

Definition 2.16 (Bipartiter Graph)
Ein Graph $G = (V, E)$ heißt bipartit, wenn sich die Knotenmenge V in zwei Mengen V_a und V_b partitionieren lässt, so dass $\forall (s,t) \in E\colon (s \in V_a \wedge t \in V_b) \vee (t \in V_a \wedge s \in V_b)$ gilt.

Durch die *Projektion* auf eine der beiden Knotenmengen kann aus einem bipartiten Graphen ein unipartiter Graph für die entsprechende Knotenmenge erstellt werden.

Definition 2.17 (Zugehöriger unipartiter Graph)
Sei $G = (V, E)$ ein bipartiter Graph mit der Knotenmenge $V = V_a \cup V_b$. Der Graph $G_a = (V_a, E_a)$ mit der Kantenmenge $E_a := \{(s,t) \mid (s,x) \in E, (x,t) \in E, s,t \in V_a, x \in V_b\}$ ist der zu G und der Knotenmenge V_a gehörig unipartite Graph.

Im zugehörigen unipartiten Graphen sind zwei Knoten also genau dann verbunden, wenn im bipartiten Graph ein (gerichteter) Weg der Länge zwei zwischen diesen beiden Knoten existiert.

Die Betrachtung von Flüssen innerhalb von Graphen ist für die Analyse von metabolischen Reaktionsnetzen (siehe Kapitel 6) erforderlich. Hierzu ist der Begriff des maximalen Flusses zwischen zwei Knoten wesentlich.

Definition 2.18 (Fluss im Graph, Wert eines s-t-Flusses, Maximaler Fluss)
Sei $G = (V, E)$ ein Graph, $c\colon E \mapsto \mathbb{R}^+$ eine Gewichtsfunktion, die Kapazitätsfunktion *zu G, und $s, t \in V$ zwei Knoten, die* Quelle *(Source) bzw. die* Senke *(Target). Eine Funktion $f\colon E \mapsto \mathbb{R}^+$ heißt* s-t-Fluss*, wenn f die beiden folgenden Bedingungen erfüllt:*

1. *Kapazitätsbeschränkung:* $\forall e \in E\colon 0 \leq f(e) \leq c(e)$
2. *Flusserhaltung:* $\forall v \in V \setminus \{s,t\}\colon \sum_{u \in \Gamma^-(v)} f((u,v)) = \sum_{w \in \Gamma^+(v)} f((v,w))$

Der Wert *eines s-t-Flusses f ist definiert als $|f| := \sum_{v \in \Gamma^+(s)} f((s,v)) - \sum_{v \in \Gamma^-(s)} f((v,s))$. Ein* maximaler Fluss *(engl.* maximum flow*)* $\mathrm{MaxFlow}(G, c, s, t)$ *von s nach t in einem Graphen G mit Kapazitätsfunktion c ist ein Fluss f mit maximalem Wert, der den obigen Bedingungen gehorcht.*

Der Wert eines maximalen Flusses wird im Folgenden mit $\mathrm{MaxFlowValue}(G, c, s, t)$ bezeichnet. Das Problem der Bestimmung des maximalen s-t-Flusses zu einem gegebenen Graph wird *Max-Flow-Problem* genannt. Für die Berechnung des maximalen Flusses existieren eine ganze Reihe von Algorithmen. Der Algorithmus von Goldberg und Tarjan benötigt beispielsweise eine Laufzeit[1] von $\mathcal{O}(|V||E| \log(|V|^2/|E|))$ [1, 64].

2.2. Zentralitätsdefinition

Im Folgenden wird der Begriff *Zentralität* definiert. Neben diesem Begriff werden die Begriffe *Zentralitätsmaß* und *Zentralitätsindex* im Folgenden synonym verwendet. Alle in den späteren Abschnitten verwendeten Zentralitätsmaße erfüllen die folgende Definition:

[1] Die Laufzeit eines Algorithmus F in Abhängigkeit von der Eingabegröße n wird in der so genannten O-Notation angegeben. Die Laufzeit von F liegt in der Komplexitätsklasse $\mathcal{O}(G)$ (geschrieben $F \in \mathcal{O}(G)$) genau dann, wenn $\exists c > 0\ \exists n_0\ \forall n > n_0 : |F(n)| \leq c \cdot |G(n)|$. Die Laufzeit von F ist also durch die Funktion G nach oben beschränkt [92].

Definition 2.19 (Zentralität)
Sei $G = (V, E)$ ein Graph bzw. (G, ω) ein gewichteter Graph. Eine Zentralität *ist eine Abbildung von der Knotenmenge V in die reellen Zahlen ($\mathcal{C}\colon V \mapsto \mathbb{R}$). Im Fall eines ungewichteten Graphen ist diese Abbildung nur abhängig von der Struktur des Graphen und für einen gewichteten Graphen ist diese Abbildung abhängig von der Graphstruktur und der Gewichtsfunktion ω.*

Durch eine Zentralität $\mathcal{C}$ können die Knoten eines Graphen verglichen werden. Es gilt dabei, dass ein Knoten v_1 „zentraler" als ein anderer Knoten v_2 ist genau dann, wenn $\mathcal{C}(v_1) > \mathcal{C}(v_2)$ ist. Eine Zentralität ist folglich geeignet, die Knoten eines Graphen zu ordnen und ermöglicht die Erstellung einer Reihenfolge der Knoten. Zentralitäten lassen sich für beide Komponenten eines Graphen, für Knoten und für Kanten, definieren [95]. Im Rahmen dieser Dissertation werden allerdings ausschließlich Knotenzentralitäten betrachtet.

Zentralitätswerte sollten immer nur im Kontext des zu analysierenden Graphen verglichen werden. Im Allgemeinen gilt nämlich, dass die Zentralitätswerte der Knoten eines Graphen bzgl. einer Zentralität $\mathcal{C}_a$ unvergleichbar mit den Zentralitätswerten bzgl. einer zweiten Zentralität $\mathcal{C}_b$ sind. Auch die Zentralitätswerte von Knoten unterschiedlicher Graphen sind bzgl. einer festen Zentralität im Allgemeinen nicht vergleichbar.

Je nach dem was für ein Prozess durch einen Graphen modelliert wird unterscheidet sich die Wichtigkeit der einzelnen Knoten. Bei einem Graphen der eine Wahl beschreibt, d.h. die Stimmvergabe zwischen verschiedenen Personen dokumentiert, ist der Knoten mit der höchsten Stimmanzahl der Gewinner der Wahl. Die Bewertung der Knoten erfolgt dabei ausschließlich lokal; es wird die unmittelbare Nachbarschaft eines Knotens für die Bewertung herangezogen. Im Fall eines Schienennetzes ist die Wichtigkeit eines Knotens hingegen von der Streckenführung durch das Gesamtnetz abhängig. Wenn beispielsweise ein Knoten als einziger zwei Teilnetze verbindet, dann ist dieser für die Funktionsweise des Schienennetzes essentiell. Die Verbindung aus dem durch den Graphen modellierten Prozess und die dazugehörige Beschreibung der „Wichtigkeit" der Knoten wird im Folgenden als *Bewertungsmaßstab* bezeichnet. Jedes Zentralitätsmaß macht Annahmen über diesen Bewertungsmaßstab und für die Auswahl einer Zentralität ist dieser von entscheidender Bedeutung. Die Wahl eines „falschen" Zentralitätsmaßes, d.h. einem Zentralitätsmaßes, das den Bewertungsmaßstab nicht widerspiegelt, führt üblicherweise zu einer Bewertung der Knoten, die nicht dem erwarteten entspricht [22].

2.3. Zentralitätsmaße

Zentralitätsmaße können anhand des von ihnen eingesetzten Bewertungsmaßstabes in Kategorien eingeteilt werden. Diese Einteilung[2] dient u.a. dazu, ein Zentralitätsmaß anhand eines gegebenen bzw. gewählten Bewertungsmaßstabes auszuwählen. Die im Folgenden beschriebenen Zentralitäten können in insgesamt fünf Kategorien eingeteilt werden:

Nachbarschafts-basiert Nur die unmittelbare Umgebung eines Knotens wird in der Berechnung des Zentralitätswertes berücksichtigt.

Weg-basiert Der Weg von einem Knoten zu den anderen Knoten innerhalb des Graphen wird bei der Berechnung des Zentralitätswertes berücksichtigt.

[2] Eine detailliertere Einteilung verschiedener Zentralitätsmaße in Kategorien wurde von Borgatti & Everett und Koschützki *et al.* vorgenommen [23, 96].

Beobachtungs-basiert Der Zentralitätswert eines Knotens hängt davon ab, auf wievielen Wegen zwischen anderen Knoten dieser liegt.

Rückkopplungs-basiert Die Zentralitätswerte der anderen Knoten beeinflussen den Zentralitätswert des betrachteten Knotens.

Anzahl deaktivierter Netzwerkelemente Der Zentralitätswert eines Knotens hängt von der Anzahl der durch die Entfernung des Knotens vom Graphen abgetrennten Knoten ab.

Alle in dieser Dissertation verwendeten Zentralitätsmaße werden nun in der Reihenfolge der obigen Konzepte vorgestellt. Dabei wird für Zentralitäten, die bereits mehrfach zur Analyse molekularbiologischer Netzwerke eingesetzt wurden bzw. die für einen Vergleich zu den in dieser Dissertation neu definierten Zentralitäten genutzt werden, die entsprechenden Definitionen angegeben. Für Zentralitäten, die bisher nur in einigen Fällen zur Analyse molekularbiologischer Netzwerke eingesetzt wurden, wird hier nur eine Beschreibung angegeben.

Es wird im Folgenden immer ein gerichteter und gewichteter Graph vorausgesetzt. Auf die Besonderheiten für ungerichtete gewichtete Graphen wird im Einzelfall eingegangen. Ungewichtete Graphen werden als Graphen mit einem konstanten Kantengewicht von 1 modelliert. Fast alle Zentralitäten werden im Folgenden mit ihren englischsprachigen Namen benannt, da diese in der Literatur weiter verbreitet sind und der Einstieg in die zitierte Literatur hierdurch vereinfacht wird. Die Laufzeitkomplexität der zur Berechnung der Zentralitäten verwendeten Algorithmen wird in diesem Kapitel nicht angegeben; statt dessen wird auf die entsprechende Literatur verwiesen [75].

2.3.1. Nachbarschafts-basierte Zentralität

Bei dieser Zentralität werden nur die direkt mit dem betrachteten Knoten verbundenen Knoten bei der Bestimmung des Zentralitätswertes berücksichtigt.

Degree-Zentralität

Die *Degree*-Zentralität weist jedem Knoten die Summe der Gewichte der Kanten, über die der betrachtete Knoten mit anderen Knoten verbunden ist, zu. Bei gerichteten Graphen werden dabei die eingehenden und die ausgehenden Kanten unterschieden. Folglich existieren für gerichtete Graphen zwei Definitionen für die *Degree*-Zentralität, die *In-Degree*-Zentralität und die *Out-Degree*-Zentralität. Bei ungerichteten Graphen unterscheiden sich die beiden Zentralitätswerte nicht. In diesem Fall wird die Zentralität einfach *Degree*-Zentralität genannt.

Definition 2.20 (*In-Degree*/*Out-Degree*-Zentralität)
Sei (G, ω) ein gerichteter, gewichteter Graph. Die In-Degree-Zentralität *ist definiert als*

$$\mathcal{C}_{ideg}(v) := \sum_{u \in \Gamma^-(v)} \omega((u, v))$$

Die Out-Degree-Zentralität *ist definiert als*

$$\mathcal{C}_{odeg}(v) := \sum_{w \in \Gamma^+(v)} \omega((v, w))$$

Die *Degree*-Zentralität ist ein lokales Zentralitätsmaß, sie berücksichtigt nur die unmittelbare Umgebung des jeweiligen Knotens. Die Struktur des Graphen über die unmittelbare Umgebung hinaus wird bei dieser Zentralität nicht berücksichtigt. Schleifen innerhalb des Graphen werden bei der Berechnung der Zentralitätswerte berücksichtigt und führen zu einer Erhöhung des Zentralitätswertes. Üblicherweise wird deshalb ein schleifenfreier Graph bei der Berechnung der *Degree*-Zentralität vorausgesetzt. Ein hohes Kantengewicht führt bei dieser Zentralität zu einem hohen Zentralitätswert.

2.3.2. Weg-basierte Zentralitätsmaße

Die in diesem Abschnitt beschriebenen Zentralitäten nutzen Informationen über Wege vom betrachteten Knoten zu den anderen Knoten innerhalb des Graphen als wesentliches Kriterium für die Bewertung des Knotens. Sie unterscheiden sich dabei einerseits in der genauen Definition des Begriffs Weg und andererseits in der Art, wie die Informationen über die Wege in die Berechnung der Zentralitätswerte einfließen.

Sechs der Zentralitäten verwenden die Distanz, d.h. die Länge eines kürzesten Weges zwischen zwei Knoten, als zugrunde liegendes Maß. Da diese zwischen Knoten, die nicht durch einen Weg verbunden sind, nicht definiert ist, wird für diese Zentralitäten üblicherweise angenommen, dass der zu analysierende Graph stark zusammenhängend ist. Schleifen werden bei der Berechnung der Distanz ignoriert, d.h. der betrachtete Graph kann Schleifen enthalten, diese haben allerdings für die Berechnung der Zentralität keine Bedeutung. Kantengewichte werden bei der Berechnung der kürzesten Wege als Entfernungen interpretiert. Folglich führt ein hohes Kantengewicht dazu, dass die betreffende Kante bei der Berechnung des kürzesten Weges weniger verwendet wird. Dieser Umstand unterscheidet sich deutlich von der *Degree*-Zentralität, da bei der *Degree*-Zentralität eine hohes Kantengewicht zu einem hohen Zentralitätswert führt. Eine Zentralität verwendet einen angenommenen Stromfluss in einem elektrischen Schaltkreis als Maß für die Entfernung. Drei der vorgestellten Zentralitäten berücksichtigen *alle* innerhalb des Graphen möglichen Wege und die letzte Zentralität kann als Erweiterung der *Degree*-Zentralitäten betrachtet werden, da bei dieser Zentralität nur „kurze“ Wege, nämlich transitive Beziehungen, gezählt werden.

Die ersten drei der im Folgenden beschriebenen Zentralitätsmaße haben ihren Ursprung in der Standortwahl bzw. Standortoptimierung (engl. *Facility Location Problems*). Hierbei wird die Frage nach der Wahl eines geeigneten Standorts für ein Gebäude, wenn zwischen mehreren möglichen Standorten zu wählen ist, beantwortet [39].

Eccentricity-Zentralität

Die *Eccentricity*-Zentralität kann sehr anschaulich anhand eines Standortoptimierungsproblems erklärt werden: Innerhalb einer Stadt soll der Standort für eine Rettungswache gewählt werden (engl. *Emergency Facility Location Problem*). Das Straßennetz wird hierzu als Graph modelliert. Bei einem Notfall am Knoten t ist somit die Entfernung $\mathrm{dist}(s,t)$ zu überwinden, um vom Standort s des Rettungswagens in der Rettungswache zur Unglücksstelle zu gelangen. Da während der Planung davon ausgegangen werden muss, dass an jedem Knoten ein Notfall eintreten kann, ist die maximal mögliche Distanz ($\mathrm{ecc}(s) = \max\{\mathrm{dist}(s,t) \colon t \in V\}$) für die Betrachtung heranzuziehen. Die besten Standorte für eine Rettungswache sind folglich die Standorte s, an denen die maximal mögliche Distanz $\mathrm{ecc}(s)$ minimiert wird.

Die zu diesem Optimierungsproblem gehörige Zentralität heißt *Eccentricity*-Zentralität (kurz *Eccentricity*). Bei dieser Zentralität ist ein Knoten umso zentraler, je weniger Schritte

notwendig sind, um von diesem Knoten zu jedem anderen Knoten zu kommen.

Definition 2.21 (*Eccentricity*-Zentralität)
Sei (G, ω) ein gerichteter, gewichteter Graph. Die Eccentricity-Zentralität *[67] ist definiert*[3] *als*

$$\mathcal{C}_{ecc}(s) := \frac{1}{\mathrm{ecc}(s)} = \frac{1}{\max\{\mathrm{dist}(s,t) \colon t \in V\}}$$

Closeness-Zentralität

Um die Nutzung eines Dienstleistungszentrums für potentielle Kunden so angenehm wie möglich zu machen, sollte dieses möglichst für alle Kunden schnell zu erreichen sein (engl. *Service Facility Location Problem*). Im Unterschied zum vorigen Problem ist also bei der Wahl des Standortes eines Dienstleistungszentrums die Summe der Entfernungen aller möglichen Kunden zu minimieren. Analog zur obigen Argumentation für die *Eccentricity*-Zentralität ist deshalb die *Closeness*-Zentralität über die Summe der Distanzen definiert:

Definition 2.22 (*Closeness*-Zentralität)
Sei (G, ω) ein gerichteter, gewichteter Graph. Die Closeness-Zentralität *[146] ist definiert als*

$$\mathcal{C}_{clo}(s) := \frac{1}{\sum_{t \in V} \mathrm{dist}(s,t)}$$

Knoten mit hohem *Closeness*-Zentralitätswert können also im Vergleich zu Knoten mit niedrigem Zentralitätswert alle anderen Knoten schneller erreichen.

Die Definitionen für *Eccentricity* und *Closeness* berechnen beide die so genannte *Out-Closeness* bzw. *Out-Eccentricity*. Analog zu dieser lässt sich durch das vertauschen der Rolle von s und t in den beiden Definitionen die *In*-Varianten definieren.

Centroid Value

Die Zentralität *Centroid Value* kann ebenfalls für die Berechnung eines optimalen Standortes eines Dienstleistungszentrums verwendet werden. Im Unterschied zur *Closeness*-Zentralität berücksichtigt diese Zentralität zusätzlich die Standorte von Konkurrenten (engl. *Competitive Facility Location Problem*). Ein Knoten erhält nur dann einen hohen Zentralitätswert, wenn dieser Knoten im Hinblick auf die Distanz zu den Kunden „besser“ erreichbar ist, als die Standorte der Konkurrenten [159].

Im Unterschied zu den vorigen Zentralitäten können die Zentralitätswerte bei der Zentralität *Centroid Value* auch negativ sein. Da diese Zentralität zur Analyse molekularbiologischen Netzwerken bisher nur in einer Publikation verwendet wurde und für Vergleiche im Folgenden nicht verwendet wird, wird diese Zentralität hier nicht im Detail vorgestellt.

Radiality und *Integration*

Die von Valente & Foreman vorgeschlagenen Zentralitäten *Radiality* und *Integration* sind ähnlich zur *Closeness*-Zentralität: alle drei Zentralitäten summieren die Distanz vom betrachteten Knoten zu den anderen Knoten im Graphen auf. Der wesentliche Unterschied

[3] Hage & Harary verwenden nicht das Reziproke, sondern direkt den Wert $\mathrm{ecc}(s)$ und definieren, dass ein zentraler Knoten einen geringen Wert hat. Um die Konsistenz mit den anderen Zentralitäten herzustellen, wird der Kehrwert verwendet.

besteht in der Definition des Distanzmaßes. Für die Zentralitäten *Radiality* und *Integration* wird die umgekehrte Distanzmatrix RD verwendet. Diese Matrix ist definiert als $RD_{ij} := \text{diameter}(G) + 1 - \text{dist}(i,j)$ wobei $\text{diameter}(G)$ den größten Distanzwert in G, d.h. $\text{diameter}(G) := \max\{\text{dist}(s,t)\colon s,t \in V\}$, bezeichnet. Für die Zentralität *Radiality* wird diese Matrix zeilenweise und für die Zentralität *Integration* spaltenweise aufsummiert:

Definition 2.23 (*Radiality* und *Integration*)
Sei (G,ω) ein gerichteter, gewichteter Graph und $n := |V|$. Die Zentralitäten Radiality und Integration [165] sind definiert als

$$\mathcal{C}_{rad}(i) := \frac{\sum_{j=1 \wedge i \neq j}^{n} RD_{ij}}{n-1} \qquad \text{und} \qquad \mathcal{C}_{int}(j) := \frac{\sum_{i=1 \wedge i \neq j}^{n} RD_{ij}}{n-1}$$

Knoten mit hohem *Radiality*-Wert können sehr schnell die anderen Knoten erreichen und Knoten mit hohem *Integration*-Wert können von anderen Knoten schnell erreicht werden. Auf Grund der Definition der umgekehrten Distanzmatrix können die beiden Zentralitäten, im Unterschied zur *Closeness*-Zentralität, auch auf nicht stark-zusammenhängende Graphen angewendet werden. Hierzu wird festgelegt, dass im Fall der Nichtexistenz eines s,t-Weges $\text{dist}(s,t) = 0$ gilt.

Load Point

Rahman & Schomburg haben eine Zentralität mit dem Namen *Load Point* vorgestellt [140]. Im Unterschied zu den vorigen über kürzesten Wegen definierten Zentralitäten werden bei dieser Zentralität *alle* kürzesten Wege (*k-shortest paths*[4] mit genügend großem k), auf denen der betrachtete Knoten liegt, berücksichtigt. Wie bei der *Degree*-Zentralität wird bei dieser Zentralität der *In-Load Point*- und der *Out-Load Point*-Wert unterschieden. Definiert ist diese Zentralität wie folgt:

Definition 2.24 (*In-Load Point*-/*Out-Load Point*-Zentralität)
Sei (G,ω) ein gerichteter, gewichteter Graph. Die In-Load Point- und die Out-Load Point-Zentralitäten[5] *[140] sind definiert als*

$$\mathcal{C}_{ilop}(v) := \sum_{s \in V(G) \wedge s \neq v} \sum_{t \in V(G)} \frac{\text{nasp}(s,t,v)}{\mathcal{C}_{ideg}(v)}$$

$$\mathcal{C}_{olop}(v) := \sum_{s \in V(G)} \sum_{t \in V(G) \wedge t \neq v} \frac{\text{nasp}(s,t,v)}{\mathcal{C}_{odeg}(v)}$$

wobei $\text{nasp}(s,t,v) := |\{P \mid P \in \text{asp}(s,t), ((v,x) \in P \vee (x,v) \in P)\}|$ *die Anzahl aller kürzesten Wege von s nach t in denen der Knoten v vorkommt und hier drin* $\text{asp}(s,t)$ *die Menge aller kürzesten Wege von s nach t bezeichnet.*

Wie die beiden Zentralitäten *Integration* und *Radiality* kann diese Zentralität auch auf nicht-zusammenhängende Graphen angewendet werden.

[4] Bei der Berechnung der k-kürzesten Pfade zwischen zwei Knoten s und t wird nicht nur ein möglicher Pfad mit kürzester Länge sondern bis zu k Pfade berechnet. Bei der Wahl eines genügend großen k werden so alle kürzesten Pfade zwischen s und t ermittelt [43].

[5] Die von Rahman & Schomburg angegebene Normierung der Zentralitätswerte wurde in diese Definition nicht übernommen.

Current-Flow Closeness (auch *Information Centrality*)

Die bisher in diesem Abschnitt vorgestellten Zentralitäten nutzen für die Berechnung der Zentralitätswerte die Länge bzw. die Anzahl der *kürzesten* Wege innerhalb des Graphen. Eine Alternative zum Modell des kürzesten Weges ist der Zufallsweg (engl. *Random-Walk*). Hierbei wird auf dem Weg vom Knoten s zum Knoten t an jedem Knoten eine zufällige Auswahl der nächsten Kante vorgenommen. Diese Auswahl folgt einer uniformen Verteilung bzw. einer Verteilung, die vom Kantengewicht abhängig ist.

Zwischen Zufallswegen in Graphen und dem elektrischem Fluss in einem Schaltkreis besteht ein sehr enger Zusammenhang [40]. Die Begriffswelt der elektrischen Schaltkreise und die gesamte Theorie zur Berechnung ihrer Eigenschaften lässt sich deshalb für die Definition von Zentralitäten nutzen. Eine der Zentralitäten, die auf der Basis von Schaltkreisen definiert ist, ist die *Current-Flow Closeness*-Zentralität [26, 161]. Sie ist definiert als die reziproke Summe der effektiven Widerstände zwischen dem Knoten s und allen anderen Knoten t. Berechnet werden kann diese Zentralität für alle ungerichteten, zusammenhängenden und gewichteten Graphen, wobei das Kantengewicht der einzelnen Kanten den jeweiligen elektrische Leitwert angibt.

Bisher wurde diese Zentralität erst einmal zur Analyse von molekularbiologischen Netzwerken eingesetzt, dabei unter dem Namen *Information Centrality*. Da sie im Folgenden nicht für Vergleiche verwendet wird, wird sie in diesem Kapitel nicht definiert.

Katz Status Index

Bei den bisher in diesem Abschnitt beschriebenen Zentralitäten hängt der Zentralitätswert der Knoten von kürzesten bzw. zufälligen Wegen innerhalb des betrachteten Graphen ab. Auf *alle* innerhalb eines Graphen vorkommenden Wege wird die Betrachtung beim *Status Index* von Katz verallgemeinert. Bei dieser Zentralität wird die Anzahl aller Wege, die zu einem betrachteten Knoten führen, korrigiert um einem von der Länge der Wege abhängigen Abschwächungsfaktor, gezählt. Dieser Abschwächungsfaktor α muss positiv sein und hat einen entscheidenden Einfluss auf die Bewertung der Knoten: Wenn α nahe Null gewählt wird, dann bewertet der *Status Index* die Knoten sehr ähnlich zum *Out-Degree* ($\mathcal{C}_{odeg}$) und je höher α gewählt wird, desto stärker ist der Einfluss der längeren Wege auf den Zentralitätswert. Nach oben ist der Abschwächungsfaktor begrenzt durch $\frac{1}{|\lambda|}$ für alle Eigenwerte λ der Adjazenzmatrix A zum untersuchten Graphen G.

Berechnet wird die Zentralität auf der Basis der Adjazenzmatrix A und der Überlegung, dass die Anzahl aller von einem Knoten j zu einem Knoten i führenden Wege aus den Potenzen der Adjazenzmatrix ablesbar ist. Es gilt nämlich, dass im Eintrag $(A^k)_{ji}$ der k-ten Potenz der Adjazenzmatrix genau die Anzahl aller am Knoten j beginnenden und am Knoten i endenden Wege der Länge k steht [162]. Der Zentralitätswert eines Knotens i ergibt sich dann als Summation über alle möglichen Wege zu i: $\mathcal{C}_{katz}(i) := \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji}$. Durch die Einschränkung des Abschwächungsfaktor α auf das Intervall $[0, \frac{1}{|\lambda|}[$, für alle λ aus den Eigenwerten von A, und ein zusätzliches Verbot von Schleifen innerhalb des Graphen ist die Konvergenz obiger Potenzreihe sichergestellt [51].

Definiert ist der *Status Index* von Katz wie folgt:

Definition 2.25 (Katz Status Index)
Sei (G, ω) ein gerichteter, gewichteter und schleifenfreier Graph und A die gewichtete Ad-

jazenzmatrix zu G und ω. Der Status Index*[87] ist definiert als*

$$\mathcal{C}_{katz}(i) := \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji}$$

Da Schleifen der Konvergenz der Potenzreihe entgegen stehen, sind diese bei dieser Zentralität explizit verboten. Hohe Kantengewichte führen beim *Status Index* zu einem höheren Zentralitätswert. Der zu untersuchende Graph muss, im Vergleich zu vielen anderen auf Wegen basierenden Zentralitätsmaßen nicht zusammenhängend sein. Die Betrachtung des inversen Graphen G_{inv} zum Graph G ermöglicht die Bewertung der Knoten anhand der ausgehenden anstatt der eingehenden Wege.

Subgraph-Centrality und *Bipartivity*

Unter dem Namen *Subgraph Centrality* wurde von Estrada & Rodríguez-Velázquez eine Zentralität vorgestellt, die die Anzahl der geschlossenen Wege (siehe Definition 2.7), die vom betrachteten Knoten ausgehen, zählt [48]. Kürzere geschlossene Wege werden hierbei mit einem höheren Gewicht als längere Wege berücksichtigt. Da der bei der Berechnung der Zentralität betrachtete Subgraph nicht variabel gewählt werden kann, wäre *Closed-Walk Centrality* eine passendere Bezeichnung für diese Zentralität.

Eine Variante der *Subgraph Centrality* ist die *Bipartivity* [47]. Bei dieser Zentralität werden geschlossene Wege mit gerader Länge und Wege mit ungerader Länge unterschieden. Knoten, von denen ausschließlich geschlossene Wege gerader Länge ausgehen erhalten den Zentralitätswert 1 und Knoten, von denen ausschließlich Wege ungerader Länge ausgehen, erhalten einen Zentralitätswert nahe 0.5, wobei die Zentralitätswerte dieser Zentralität auf das Intervall $]0.5, 1]$ beschränkt sind. Ein höherer Zentralitätswert bedeutet somit, dass der Knoten eher Ausgangspunkt von geschlossenen Wegen mit gerader Länge ist. Da in bipartiten Graphen ausschließlich geschlossene Wege mit gerader Länge auftreten [38], werden Knoten mit hohem Zentralitätswert bzgl. der obigen Erklärung als „bipartite Knoten“ bezeichnet.

Bisher wurden die beiden Zentralitäten erst in wenigen Fällen zur Analyse von molekularbiologischen Netzwerken eingesetzt. Da beide Zentralitäten in dieser Dissertation auch nicht für Vergleiche verwendet werden, werden sie hier nicht im Detail beschrieben.

Transitivity

In der von Wuchty beschriebenen Zentralität *Transitivity* werden die Anzahl der transitiven Beziehungen, in denen der betrachtete Knoten beteiligt ist, gezählt [172]. Hierzu wird für jedes Knotentripel (v_1, v_2, v_3) geprüft, ob je eine Kante zwischen v_1 und v_2 und zwischen v_2 und v_3 existiert. Im Fall, dass beide Kanten existieren wird geprüft, ob auch eine Kante zwischen v_1 und v_3 existiert. Ist dieses der Fall, dann wird der Transitivitätswert von v_1 um eins erhöht. Nachdem alle möglichen transitiven Beziehungen des Knotens v_1 gezählt wurden, wird der relative Wert, d.h. die Anzahl der tatsächlichen transitiven Beziehungen gegenüber der möglichen ermittelt. Dieser Wert ist der Transitivitätswert des Knotens v_1. Auch die *Transitivity* wird im Folgenden nicht für Vergleiche verwendet und folglich hier nicht explizit definiert.

2.3.3. Beobachtungs-basierte Zentralitätsmaße

Grundidee der Beobachtungs-basierten Zentralitätsmaße ist, dass ein Knoten der Bestandteil eines Weges von einem Knoten zu einem anderen Knoten ist, eine „Kommunikation“ zwischen diesen beiden Knoten *beobachten* könnte. Dabei ist ein Knoten dann als besonders wichtig anzusehen, wenn eine große Anzahl von Wegen über ihn verlaufen. Im Unterschied zu den Weg-basierten Zentralitäten ist der betrachtete Knoten nicht Start- bzw. Endpunkt des Weges sondern einer der Knoten *auf dem Weg*.

Vergleichbar zu den Weg-basierten Zentralitätsmaßen können die Beobachtungs-basierten Maße ebenfalls über verschiedene Arten von Wegen definiert werden. Im Folgenden werden zwei Arten von Wegen, kürzeste Wege und Zufallswege, betrachtet.

Shortest-Path Betweenness

Die erste der beiden Beobachtungs-basierten Zentralitätsmaße nutzt kürzeste Wege (engl. *shortest paths*) für die Bewertung der Knoten. Die Anzahl aller kürzesten Wege zwischen den Knoten s und t wird im Folgenden mit σ_{st} bezeichnet und die Anzahl der kürzesten Wege zwischen s und t, auf denen der Knoten v vorkommt, mit $\sigma_{st}(v)$. Dabei wird ein eventuell vorliegendes Kantengewicht bei der Bestimmung der kürzesten Wege berücksichtigt. Das Verhältnis der Anzahl der kürzesten s,t-Wege die den Knoten v nutzen, gegenüber der Anzahl aller kürzesten s,t-Wege ist $\delta_{st}(v) := \frac{\sigma_{st}(v)}{\sigma_{st}}$. Falls kein Weg von s nach t existiert, dann sei $\delta_{st}(v) := 0$. Der Wert $\delta_{st}(v)$ liegt somit immer zwischen 0 und 1 und gibt die Wahrscheinlichkeit an, mit der der Knoten v bei einer „Kommunikation“ zwischen s und t genutzt wird.

Die Zentralität *Shortest-Path Betweenness* bewertet die Knoten anhand dieser Wahrscheinlichkeit $\delta_{st}(v)$:

Definition 2.26 (*Shortest-Path Betweenness*-Zentralität)
Sei (G, ω) ein gerichteter, gewichteter Graph. Die Shortest-Path Betweenness-Zentralität *[11, 57] ist definiert als*

$$\mathcal{C}_{spb}(v) = \sum_{(s \in V \wedge s \neq v)} \sum_{(t \in V \wedge t \neq v)} \delta_{st}(v)$$

Ein Knoten erhält dann einen hohen Zentralitätswert, wenn er für viele „Kommunikationen“ verwendet werden muss. Die *Shortest-Path Betweenness*-Zentralität kann auch auf nicht stark zusammenhängende Graphen angewendet werden. Schleifen werden bei der Berechnung der Zentralitätswerte nicht berücksichtigt und ein höheres Kantengewicht führt bei den mit der Kante verbundenen Knoten zu einer Verringerung der Zentralitätswerte.

Current-Flow Betweenness (auch *Random-Walk Betweenness*)

Ähnlich zur Weg-basierten Zentralität *Current-Flow Closeness* lässt sich auch eine Beobachtungs-basierte Zentralität für elektrische Schaltkreise definieren. Hierbei wird der Anteil des elektrischen Flusses an einem Knoten v betrachtet, der bei einem Stromfluss von einem Knoten s zu einem anderen Knoten t über v fließt. Wie auch die *Current-Flow Closeness* kann die *Current-Flow Betweenness* nur auf ungerichtete, zusammenhängende und gewichtete Graphen angewendet werden. Die von Newman definierte Zentralität *Random-Walk Betweenness* und die *Current-Flow Betweenness* Zentralität sind äquivalent [26, 127].

Bisher wurde diese Zentralität erst einmal zur Analyse von molekularbiologischen Netzwerken eingesetzt, dabei unter dem Name *Random-Walk Betweenness*. Da sie im Folgenden nicht für Vergleiche verwendet wird, wird sie in diesem Kapitel nicht definiert.

2.3.4. Rückkopplungs-basierte Zentralitätsmaße

Fast alle bisher betrachteten Zentralitäten sind über Wege innerhalb des Graphen definiert. Im Folgenden werden drei Zentralitäten vorgestellt, die über Rückkopplungsprozesse definiert sind. Bei diesen Zentralitäten hängt der Zentralitätswert eines Knotens von den Zentralitätswerten einiger oder aller anderen Knoten ab. Diese Betrachtung führt üblicherweise zu einem linearen Gleichungssystem, welches mit den Methoden der linearen Algebra gelöst werden kann. Häufig werden die Zentralitätswerte der Rückkopplungs-basierten Zentralitäten deshalb als Vektoren (geschrieben $\vec{\mathcal{C}}$) dargestellt.

Eigenvector-Zentralität

Bei der *Eigenvector*-Zentralität wird das Konzept der Rückkopplung unmittelbar sichtbar. Bei dieser Zentralität wird angenommen, dass der Zentralitätswert eines Knotens v_i sich aus der Summe der Zentralitätswerte aller Nachbarknoten bestimmt. Diese Idee führt zu folgendem linearen Gleichungssystem:

$$\begin{aligned} \mathcal{C}_{eiv}(v_1) &= \sum_{j=1}^{n} a_{1j}\mathcal{C}_{eiv}(v_j) \\ &\vdots \\ \mathcal{C}_{eiv}(v_n) &= \sum_{j=1}^{n} a_{nj}\mathcal{C}_{eiv}(v_j) \end{aligned}$$

Hierbei bezeichnen die a_{ij} die Komponenten der Adjazenzmatrix A zum betrachteten Graphen G. Dieses Gleichungssystem ($\vec{\mathcal{C}_{eiv}} = A\vec{\mathcal{C}_{eiv}}$) ist nur lösbar, wenn $\det(A - I) = 0$ gilt. Als Eigensystem ($A\vec{\mathcal{C}_{eiv}} = \lambda\vec{\mathcal{C}_{eiv}}$) ist es hingegen für beliebige reell wertige Matrizen lösbar [12]. Im Rahmen der Zentralitätsanalyse von Graphen ist die Einschränkung auf positive, reell wertige Einträge möglich[6]. Aus einer Erweiterung des Satzes von Perron-Frobenius über die Beschaffenheit der Eigenwerte und Eigenvektoren von reell wertigen Matrizen lässt sich dann eine Aussage über die Beschaffenheit eines Eigenvektors für die Matrix A ableiten: wenn der Graph G zur Adjazenzmatrix A stark zusammenhängend ist und die Kantengewichte von G nur positiv (inkl. 0) und reell wertig sind, dann existiert ein Eigenwert λ_1 der einfach ist, der betragsmäßig nicht kleiner ist, als alle anderen Eigenwerte ($\lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$) und der zu diesem Eigenwert gehörende Eigenvektor besteht aus Komponenten, die entweder alle positiv oder alle negativ sind [63]. Für die *Eigenvector*-Zentralität wird genau dieser Eigenvektor verwendet.

Definition 2.27 (*Eigenvector*-Zentralität)
Sei (G, ω) ein gerichteter, gewichteter, stark zusammenhängender Graph. Die Eigenvector-Zentralität [21] ist definiert als der Eigenvektor $\vec{\mathcal{C}_{eiv}}$ zum größten Eigenwert λ_1 des Glei-

[6] Die Kantengewichte des Graphen dürfen in diesem Fall nur positiv sein.

chungssystems $A\vec{\mathcal{C}_{eiv}} = \lambda\vec{\mathcal{C}_{eiv}}$*, bei dem entweder alle Komponenten positiv oder alle Komponenten negativ sind.*

PageRank

Die Zentralität *PageRank* wird von der Suchmaschine Google zum Bewerten von Webseiten eingesetzt [131]. Das Verfahren geht von der Annahme aus, dass ein Besucher die Webseiten wie ein *Random Surfer* besucht. Dieser startet bei einer beliebigen Seite und wählt aus den verfügbaren Links zufällig (und gleichverteilt) einen aus. Dieser Link führt auf eine neue Seite und der Besucher startet seine Auswahl erneut. Zusätzlich wird dem Besucher mit einer vorgegebenen Wahrscheinlichkeit erlaubt, durch die direkte Eingabe einer Webadresse, eine beliebige Webseite auszuwählen. Nach vielen Durchläufen lässt sich so für jede Seite eine Besuchswahrscheinlichkeit ermitteln. Je höher diese Wahrscheinlichkeit ist, desto wichtiger ist diese Seite.

Mathematisch lässt sich das Verhalten des *Random Surfers* durch einen Rückkopplungsprozess beschreiben und die resultierende Besuchswahrscheinlichkeit für die einzelnen Webseiten als Zentralitätswerte für die entsprechenden Knoten interpretieren.

Definition 2.28 (PageRank)
Sei (G, ω) *ein gerichteter, gewichteter Graph. Die Zentralität* PageRank *[131] ist definiert als Lösung des folgenden linearen Gleichungssystems*

$$\vec{\mathcal{C}_{pra}} = dP\vec{\mathcal{C}_{pra}} + (1-d)1_n$$

wobei P *die Matrix* $p_{ij} := \frac{a_{ji}}{\deg^+(j)}$*,* $d \in [0\,,1[$ *und* 1_n *der* n*-Dimensionale* 1*-Vektor ist.*

Der Dämpfungsfaktor d bestimmt die Stärke des Einflusses der Graphstruktur auf die Zentralitätswerte. Ein Wert nahe 1 führt zu einer starken und ein Wert nahe 0 nur zu einer eher geringen Berücksichtigung derselben. Im Zusammenhang mit der Suchmaschine Google wird häufig ein Wert von 0,85 angegeben [102].

Für die Zentralität *PageRank* gelten folgende Einschränkungen bzw. Vorbedingungen: Der zu analysierende Graph muss nicht zwingend zusammenhängend sein. Schleifen können vorhanden sein und ändern die dazugehörigen Zentralitätswerte. Kantengewichte können bei der Berechnung berücksichtigt werden. Es gilt allerdings, dass diese nur eine „lokale“ Auswirkung haben, da diese nur in die Matrix P einfließen. Wie auch für den *Status Index* kann *PageRank* sowohl auf G als auch auf G_{inv}, den dazugehörigen inversen Graphen, angewendet werden.

Entropy

Die Kennzahl *Network Entropy* zur Beschreibung der Topologie eines Graphen wurde von Demetrius & Manke vorgeschlagen [37]. Der Anteil eines einzelnen Knotens an dieser Kennzahl erfüllt dabei die Eigenschaft einer Zentralität gemäß Definition 2.19. Auch diese Zentralität wurde bisher erst einmal zur Analyse molekularbiologischer Netzwerke eingesetzt und sie wird für weitere Vergleiche in dieser Dissertation nicht verwendet. Eine Definition wird folglich nicht angegeben.

2.3.5. Anzahl deaktivierter Netzwerkelemente

Die letzten beiden in diesem Kapitel vorzustellenden Zentralitäten basieren auf der Idee, dass beim Ausfall bzw. Entfernen eines Knotens möglicherweise Teile eines Graphen nicht

mehr erreichbar sind und damit die durch den Graphen modellierten Prozesse nicht mehr erfolgreich ablaufen können. Die Wichtigkeit eines Knotens misst sich in diesem Fall an der Größe des durch den Ausfall des Knotens verursachten Schadens.

Damage

Wenn aus einem stark zusammenhängenden[7] Graphen $G = (V, E)$ ein Knoten v entfernt wird, dann verbleibt ein Graph $G' = (V', E')$ mit $V' := V \setminus \{v\}$ und $E' := E \setminus \{(s,t) \mid s \in V, t \in V, s = v \vee t = v\}$ der nicht notwendigerweise stark zusammenhängend ist. Innerhalb dieses Graphen G' existiert mindestens eine starke Zusammenhangskomponente[8] $G'' \subseteq G'$. Die Größe dieser Zusammenhangskomponente ist dabei durch $|V(G)| - 1$ nach oben beschränkt, da entweder nur der Knoten v vom Graphen G abgetrennt wurde oder noch weitere Knoten nicht mehr mit der verbleibenden Zusammenhangskomponente verbunden sind. Die Differenz zwischen der Größe des ursprünglichen Graphen G und der Größe der Zusammenhangskomponente G'' nach dem Entfernen von v erfüllt dabei die Bedingung an eine Zentralität. Von Schmith *et al.* wurde sie *Damage*-Zentralität genannt.

Definition 2.29 (*Damage*-Zentralität)
Sei $G = (V, E)$ ein stark zusammenhängender Graph und bezeichne $\mathrm{GSC}_G(v)$ eine starke Zusammenhangkomponente von G nachdem der Knoten v entfernt wurde. Die Zentralität Damage *[151] ist dann definiert als*

$$\mathcal{C}_{dam}(v) := |V(G)| - |V(\mathrm{GSC}_G(v))|$$

Angewendet werden kann diese Zentralität auf jeden stark zusammenhängenden Graphen. Das Vorhandensein von Schleifen hat keinen Einfluss auf die Zentralitätswerte und Kantengewichte wurden in der von Schmith *et al.* angegebenen Variante nicht berücksichtigt.

Avalanche

Ein Produktionsprozess ist einen Vorgang, bei dem auf der Basis von Eingangsmaterialien Ausgangsmaterialien erstellt werden. Ein Prozess dieser Art kann als gerichteter, bipartiter Graph beschrieben werden. Hierbei repräsentiert die eine Knotenmenge die Materialien, die andere Knotenmenge die Prozesse und Kanten zwischen den Materialien und den Prozessen beschreiben die Beziehungen der Materialien und Prozesse zueinander. Da die Ausgangsmaterialien eines Prozesses gleichzeitig die Eingangsmaterialien für andere Prozesse sein können, entsteht ein Graph, der die Abhängigkeiten der Prozesse und Materialien voneinander beschreibt.

Die Deaktivierung eines Prozesses innerhalb dieses bipartiten Graphen führt dazu, dass Materialien, die exklusiv durch diesen Prozess erstellt werden, anderen Prozessen nicht mehr zur Verfügung stehen. Hierdurch können weitere Produktionsprozesse blockiert werden. Die Anzahl der nach der Deaktivierung eines Prozesses nicht mehr produzierten Materialien bzw. blockierten anderen Prozesse erfüllt dabei die Bedingungen an eine Zentralität.

Für metabolische Reaktionsnetze (siehe Abschnitt 3.4.2) wurde diese Zentralität in zwei Varianten vorgestellt: In der ersten Variante wurde die Anzahl der nicht mehr produzierten

[7] Die vorgestellte Idee ist einfach auf schwachen Zusammenhang bzw. auf Zusammenhang in ungerichteten Graphen übertragbar.
[8] Laut Definition 2.13 sind Zusammenhangskomponenten immer auch maximal.

Metaboliten (Materialien) gezählt. In diesem Fall wurde die Zentralität als *Damage* bezeichnet [104]. Die zweite Variante wurde *Avalanche* (deutsch: Lawine) genannt [62]. Bei dieser Variante wurde die Anzahl der blockierten Reaktionen (Prozesse) als Maßstab verwendet.

2.4. Zusammenfassung

In diesem Kapitel wurden Graphen und Zentralitäten definiert und es wurden insgesamt 18 Zentralitäten, eingeteilt in fünf Kategorien, vorgestellt. Alle beschriebenen Zentralitäten wurden bereits für die Analyse von molekularbiologischen Netzwerken eingesetzt bzw. werden in den folgenden Kapiteln für Vergleiche verwendet.

In der Tabelle 2.1 ist der Einfluss der drei wichtigsten Grapheigenschaften auf die „Funktionsweise“ der beschriebenen Zentralitätsmaße zusammengefasst.

Zentralität	Schleifen	Zusammenhang	Kantengewicht	Referenz
In-/Out-Degree	Erhöhen Wert	Nicht erforderlich	Hohes Gewicht/hoher Wert	
Eccentricity	Erlaubt, ignoriert	Starker zshg.	Hohes Gewicht/niedriger Wert	[67]
Closeness	Erlaubt, ignoriert	Starker zshg.	Hohes Gewicht/niedriger Wert	[146]
Centroid Value	Erlaubt, ignoriert	Starker zshg.	Hohes Gewicht/niedriger Wert	[159]
Radiality&Integration	Erlaubt, ignoriert	Nicht erforderlich	Hohes Gewicht/niedriger Wert	[165]
Load Points	Erlaubt, ignoriert	Nicht erforderlich	Hohes Gewicht/niedriger Wert	[140]
Current-Flow Closeness[a]	Erlaubt, ignoriert	Zshg.	Hohes Gewicht/hoher Wert	[26]
Katz Status Index	Nicht erlaubt	Nicht erforderlich	Hohes Gewicht/hoher Wert	[87]
Subgraph-Centrality[a]	Erlaubt, ignoriert	Zshg.	Nicht berücksichtigt	[48]
Bipartivity[a]	Erlaubt, ignoriert	Zshg.	Nicht berücksichtigt	[47]
Transitivity	Erlaubt, ignoriert	Nicht erforderlich	Nicht berücksichtigt	[172]
Shortest-Path Betweenness	Erlaubt, ignoriert	Nicht erforderlich	Hohes Gewicht/niedriger Wert	[57]
Current-Flow Betweenness[a]	Erlaubt, ignoriert	Zshg.	Hohes Gewicht/hoher Wert	[26]
Eigenvector-Zentralität	Erlaubt	Starker zshg.	Hohes Gewicht/hoher Wert	[21]
PageRank	Erlaubt	Nicht erforderlich	Hohes Gewicht/hoher Wert	[131]
Entropy[a]	Erlaubt[b]	Nicht bekannt[c]	Nicht bekannt[d]	[37]
Damage	Erlaubt, ignoriert	Starker zshg.	Nicht berücksichtigt	[151]
Avalanche[e]	Nicht erlaubt	Nicht erforderlich	Nicht berücksichtigt	[104, 62]

[a] Die Zentralität ist z.Z. nur für ungerichtete Graphen definiert.
[b] Schleifen sind laut Definition erlaubt, die Auswirkung auf den Zentralitätswert ist allerdings nicht beschrieben.
[c] In der entsprechenden Publikation wird keine Aussage hierzu gemacht.
[d] Die Auswirkung des Kantengewichtes auf den Zentralitätswert ist nicht beschrieben.
[e] Die Zentralität ist nur für bipartite Graphen definiert.

Tabelle 2.1.: Der Einfluss der drei wichtigsten Grapheigenschaften auf die „Funktionsweise“ der beschriebenen Zentralitäten. Die zweite Spalte beschreibt die Auswirkung von Schleifen auf die Zentralitätswerte. Ob und wenn ja wie der Graph zusammenhängend sein muss, ist in der dritten Spalte angegeben. Die Auswirkung eines eventuell vorhandenen Kantengewichtes auf die Zentralitätswerte ist in der vierten Spalte zusammengefasst.

3. Molekularbiologische Netzwerke und deren Grundlagen

Aufgabe der Molekularbiologie ist die Beschreibung der rund um die DNA und RNA ablaufenden Prozesse. Hierzu gehören einerseits alle Abläufe, in denen die DNA bzw. RNA unmittelbar beteiligt ist und andererseits die Vorgänge, in denen deren Produkte (im wesentlichen Proteine) genutzt werden.

Molekularbiologische Prozesse, die sich gut einzeln beschreiben und untersuchen lassen, sind in der Natur allerdings die Ausnahme. Unter normalen Umständen treten vielfältige Wechselwirkungen zwischen diesen Prozessen auf. Ein einzelner Vorgang ist somit eigentlich Bestandteil eines ganzen Netzwerkes von Prozessen, die häufig gleichzeitig mit dem zu untersuchenden Prozess auftreten und diesen beeinflussen können.

In diesem Kapitel werden die Grundlagen zum Verständnis molekularbiologischer Netzwerke vorgestellt. Zu diesen Grundlagen gehören einerseits die Umgebung, in der molekularbiologische Vorgänge ablaufen, und andererseits die Beschreibung der Vorgänge an sich. Im Folgenden wird deshalb zuerst die Zelle beschrieben und dann die einzelnen „Bausteine“ und molekularbiologischen Prozesse vorgestellt. Darauf aufbauend werden die drei betrachteten molekularbiologischen Netzwerke (Genregulationsnetze, metabolische Reaktionsnetze und Protein-Interaktionsnetze) erklärt. Signal-Transduktionsnetze, d.h. Netzwerke, die die Reizübermittlung beschreiben, werden in diesem Kapitel nicht fokussiert, da zur Zeit nur wenige Netzwerke signifikanter Größe existieren und bisher erst eine Zentralität zur Analyse eines Netzes dieses Typs eingesetzt wurde [105].

Die Darstellung in diesem Abschnitt ist größtenteils vereinfachend. Viele Spezialfälle, die in der Molekularbiologie häufig eine Rolle spielen, werden im Folgenden nicht berücksichtigt. Diese sind für das Verständnis der Methoden der Zentralitätsanalyse von molekularbiologischen Netzwerken nicht von Bedeutung. Einige der Zeichnungen in diesem Abschnitt sind aus dem „Talking Glossary of Genetics“ [74] des National Human Genome Research Institute der USA entnommen. Eine Zusammenstellung der verwendeten Literatur findet sich am Ende des Kapitels.

3.1. Umgebung des Lebens

In der Natur existieren einzellige und mehrzellige Lebewesen. Mit Blick auf einzellige Organismen ist die *Zelle* das kleinste lebensfähige System. Mehrzellige Organismen bestehen aus gleichen oder unterschiedlichen Zellen. Jede Zelle eines mehrzelligen Organismus trägt in der Regel zum Überleben des Gesamtorganismus bei und kann zudem auf eine Aufgabe spezialisiert sein.

Zellen bestehen aus einer Vielzahl von *Zellorganellen* bzw. *Kompartimenten*. Hierzu gehören beispielsweise die *Mitochondrien*, in denen biochemische Reaktionen Energie für die Zelle „herstellen“, oder die *Plastiden*, spezifische Zellbestandteile, die in der Photosynthese eine Bedeutung haben.

Abbildung 3.1.: Auf den Chromosomen befinden sich Abschnitte, die Gene kodieren. Bildquelle: [74]

Der *Zellkern* ist für die Molekularbiologie eines der wichtigsten Kompartimente. Im Zellkern befindet sich die Desoxyribonukleinsäure (DNA), ein Molekül, das die Erbinformationen (das Genom) speichert. Organismen, bei denen die Zellen einen Zellkern haben und bei denen die Organellen und der Zellkern von einer Membran umschlossen sind werden *Eukaryoten* genannt. Hierzu gehören im Wesentlichen alle „höheren“ Lebewesen, z.B. Pflanzen, Tiere und auch Pilze. *Saccharomyces cerevisiae*, die Bäckerhefe, ist ein in der Molekularbiologie häufig als Modellorganismus verwendeter Eukaryot. Organismen ohne Membran-umschlossenen Zellkern und ohne Organellen werden *Prokaryoten* bezeichnet. Das Darmbakterium *Escherichia coli* ist ein Prokaryot und einer der Modellorganismen der molekularbiologischen Forschung. Im Gegensatz zu den Eukaryoten befindet sich bei den Prokaryoten die DNA nicht in einem Membran-umschlossenen Zellkern sondern kann sich frei innerhalb der Zelle bewegen. Im Folgenden erfolgt eine Unterscheidung zwischen Pro- und Eukaryoten nur, falls hierfür eine Notwendigkeit besteht.

3.2. Bausteine des Lebens

Die in molekularbiologischen Prozessen auftretenden Bausteine, insbesondere Gene, Proteine und Metaboliten, werden im folgenden Abschnitt erläutert.

3.2.1. Gene und Desoxyribonukleinsäure

Die gesamte Erbinformation eines Lebewesens wird als *Genom* bezeichnet. Gespeichert ist diese Erbinformation innerhalb des Lebewesens in einer Gruppe von Molekülen, der *Desoxyribonukleinsäure* (DNS bzw. engl. *DNA*). Je nach Organismus kann die Erbinformation in mehreren sehr langen Molekülen, genannt *Chromosomen*, gespeichert sein. Der Mensch beispielsweise hat insgesamt 46 Chromosomen, die Pflanze *Arabidopsis thaliana* (Acker-Schmalwand) hat fünf Chromosomen und das Bakterium *Escherichia coli* hat ein Chromosom, welches in Form eines Ringes angeordnet ist. Die Abbildung 3.1 zeigt ein Chromosom und ein Gen, ein Teilstück der DNA.

Jedes Chromosom ist (vereinfacht dargestellt) eine Kette, die aus Kombinationen von vier Bausteinen, den *Nukleotiden*, besteht. Jedes Nukleotid wiederum besteht aus drei Bestandteilen: einer Phosphatgruppe, einem Zuckermolekül mit fünf Kohlenstoffatomen (Desoxyribose) und einer stickstoffhaltigen Base. Die Phosphatgruppe und das Zuckermolekül bilden das Rückgrat des DNA-Strangs. Hierbei sind immer zwei Nukleotide über eine sogenannte Phosphodiesterbindung miteinander verbunden. Die vier Nukleotide unterscheiden sich in

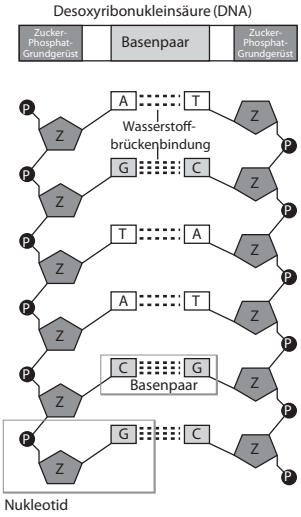


Abbildung 3.2.: Die DNA ist ein langes Molekül, bestehend aus Nukleotiden. Die Verbindungen der Phosphat- und Zuckermoleküle (P bzw. Z) bilden das Rückgrat des Moleküls. Die Wasserstoffbrückenbindungen zwischen den stickstoffhaltigen Basen führt zu einer Stabilisierung des Moleküls. Bildquelle: [74]

ihrer stickstoffhaltigen Base. Diese vier Basen sind Adenin, Guanin, Cytosin und Thymin, abgekürzt als A, G, C bzw. T. Grob abstrahiert lässt sich somit das Genom als Folge von Buchstaben aus dem Alphabet {A,G,C,T} ansehen.

DNA-Moleküle treten normalerweise paarweise auf. Die beiden Stränge der DNA sind dabei antiparallel ausgerichtet, d.h. das sogenannte $3'$-Ende mit dem Zuckermolekül steht dem sogenannten $5'$-Ende mit der Phosphatgruppe des anderen Moleküls gegenüber. Der antiparallele Strang ist eindeutig bestimmt, da zu jeder Base aus dem einen Strang die entsprechende Base auf den anderen Strang nach folgender Regel festgelegt ist: das Komplement zu Thymin ist Adenin und das Komplement zu Guanin ist Cytosin. Zusammen werden die beiden sich gegenüberstehenden Basen als *Basenpaar* bezeichnet. Aufgrund der Struktur der Moleküle bilden sich Wasserstoffbrücken zwischen den Basenpaaren aus und das gesamte Molekül ordnet sich schraubenförmig an (*Doppelhelix*). Die Abbildung 3.2 zeigt die komplementäre Anordnung der Basen in der Mitte. Jeweils links und rechts ist das Rückgrat bestehend aus dem Zuckermolekül und der Phosphatgruppe sichtbar.

Die Größe der gesamten Erbinformationen eines Organismus wird in Basenpaaren angegeben. Für *Escherichia coli* sind dieses beispielsweise $4,6 \times 10^6$ Basenpaare, die Bäckerhefe (*Saccharomyces cerevisiae*) hat ca. $12,5 \times 10^6$ Basenpaare, der Mensch (*Homo sapiens*) ca. $2,9 \times 10^9$ Basenpaare und Weizen (*Triticum aestivum*) ca. 16×10^9 Basenpaare.

Teilstücke eines DNA-Moleküls, die in RNA (siehe folgender Abschnitt) umgesetzt werden können und für den Organismus eine funktionelle Rolle spielen, werden als *Gene* bezeichnet. Zusätzlich zu den Genen existieren innerhalb des DNA-Moleküls auch Bereiche, denen keine Funktion zugeordnet werden kann; diese werden *nicht-kodierende Bereiche* genannt.

Gene werden üblicherweise durch ihre *Nukleotid-Sequenz*, d.h. die Reihenfolge der Nukleotide die in diesem DNA-Abschnitt auftreten, beschrieben. Viele schon sequenzierte Gene, deren Nukleotid-Sequenz also bekannt ist, sind in Datenbanken, beispielsweise der EMBL Nucleotide Sequence Database [101], hinterlegt.

Wie die Anzahl der Basenpaare ist auch die Anzahl der Gene zwischen den Organismen

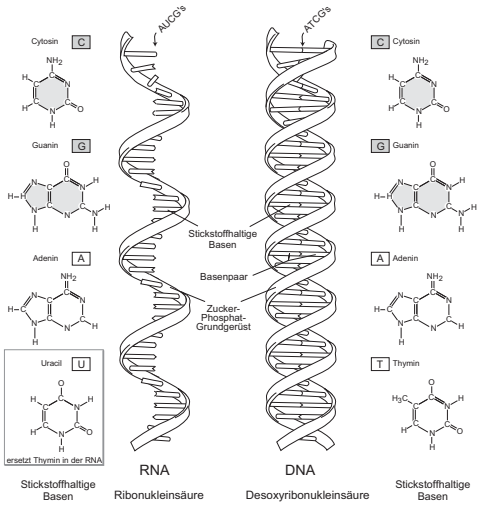


Abbildung 3.3.: Die Strukturformeln der Nukleotide und der Aufbau der RNA und der DNA im direkten Vergleich. Bildquelle: [74]

unterschiedlich. *Escherichia coli* hat beispielsweise 4 290 bekannte Gene, *Saccharomyces cerevisiae* 6 186, *Homo sapiens* ca. 30 000 und für *Triticum aestivum* geht man aktuell von ca. 90 000 Genen aus.

3.2.2. Ribonukleinsäure

Zur Verarbeitung der Erbinformationen bedient sich die Zelle einer Variante von Nukleinsäure-Molekülen, der *Ribonukleinsäure* (*RNA*). Diese Moleküle sind ähnlich aufgebaut wie die DNA. RNA und DNA unterscheiden sich nur in einigen Details: das als Rückgrat dienende Zuckermolekül ist eine Ribose anstatt einer Desoxyribose, an die Stelle der Base Thymin (T) tritt die Base Uracil (U) und, im deutlichen Gegensatz zur DNA, existiert bei der RNA kein Komplementärstrang. Die Abbildung 3.3 zeigt die Strukturformeln für die Nukleotide von RNA und DNA und den Aufbau der beiden Moleküle.

Die RNA tritt innerhalb der Zellen in verschiedenen Varianten auf. Für die Synthese von Proteinen sind die *Transfer-RNA* (tRNA) und die *Boten-RNA* (engl. *messenger-RNA*, mRNA) notwendig. Beide Varianten werden im Abschnitt über die Proteinsynthese (Abschnitt 3.3.1) genauer erklärt.

3.2.3. Proteine und Enzyme

Proteine sind die Grundbausteine der Zellen. Wie auch die DNA ist ein Protein ein Polymer bestehend aus unterschiedlichen Monomeren. Im Unterschied zur DNA bestehen Proteine aus *Aminosäuren*, d.h. aus Molekülen in denen ein Kohlenstoffatom von einem Wasserstoffatom, einer Aminogruppe (NH_2), einer Carboxylgruppe (COOH) und einem weiteren, variablen „Rest“ umgeben ist. Für den Aufbau von Proteinen stehen 20 unterschiedliche Aminosäuren zur Verfügung. Über so genannte Peptidbindungen sind die Aminosäuren miteinander zu langen Polypeptidketten verbunden. Abbildung 3.4 zeigt die allgemeine Strukturformel für Aminosäuren und die als primäre Struktur eines Proteins bezeichnete Kette von Aminosäuren. Die Primärstruktur eines Proteins wird auch *Aminosäuresequenz* oder

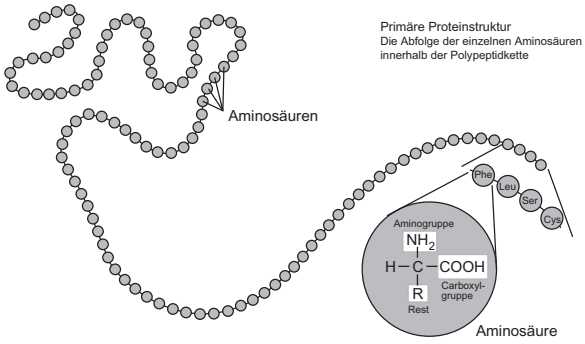


Abbildung 3.4.: Die Strukturformel für Aminosäuren und die Primärstruktur eines Proteins. Bildquelle: [74]

Proteinsequenz genannt. Auch diese Sequenzen sind über Datenbanken, beispielsweise UniProt [164], zugänglich.

Damit ein Protein seine Funktion erfüllen kann, muss es in eine dreidimensionale Struktur gefaltet sein. Dieses erfolgt aufgrund der unterschiedlichen elektrischen Ladungen der Aminosäuren und unter der Beteiligung von Chaperonen, speziellen Proteinen, die anderen Proteinen bei der Faltung „helfen“. Teilweise sind mehrere Polypeptidketten in ihrer entsprechenden dreidimensionalen Struktur notwendig um ein funktionsfähiges Protein zu bilden.

Anhand ihrer wesentlichen Funktion können Proteine in Klassen eingeteilt werden. Die für diese Dissertation wichtigen Proteine gehören zu den Enzymen und den Regulatorproteinen. *Enzyme* katalysieren biochemische Reaktionen, d.h. sie beschleunigen diese und viele Reaktionen können erst bei Vorhandensein des zugehörigen Enzyms in der notwendigen Geschwindigkeit ablaufen. *Regulatorproteinen* aktivieren oder deaktivieren andere Gene während der Genexpression (siehe Abschnitt 3.3.1).

3.2.4. Metaboliten

Moleküle, die im Rahmen des Stoffwechsels verändert werden, werden unter dem Sammelbegriff *Metabolit* zusammengefasst. Einfache Beispiele für Metaboliten sind Glukose (Traubenzucker), ein wichtiger Energielieferant der Zellen, und die Zitronensäure, die eine wichtige Rolle im Abbau von Fetten, Zuckern und Aminosäuren spielt.

3.3. Prozesse des Lebens

Basierend auf den einzelnen Bausteinen werden in diesem Abschnitt die in der Zelle ablaufenden Vorgänge, die Prozesse, beschrieben.

3.3.1. Genexpression und Proteinsynthese

Die Erstellung der Proteine, genannt *Proteinsynthese* oder *Genexpression*, erfolgt (stark vereinfacht dargestellt) in zwei Schritten. Im ersten Schritt wird ein Abschnitt der DNA „ausgelesen“ (Transkription) und in einem zweiten Schritte wird die Kette von Nukleinsäuren in eine Kette von Aminosäuren „übersetzt“ (Translation).

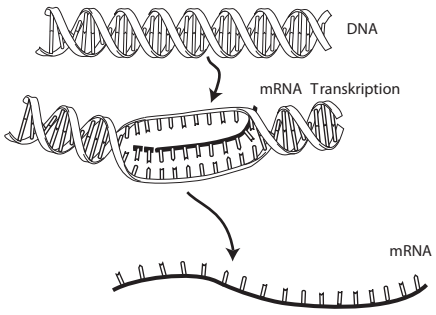


Abbildung 3.5.: Während der Transkription wird der DNA-Strang teilweise aufgetrennt und es bildet sich an einem der beiden Stränge ein komplementäres RNA-Molekül. Dieses neue Molekül löst sich von der DNA und steht dann als mRNA zur Verfügung. Bildquelle: [74]

Die *Transkription*, d.h. das Auslesen eines Abschnitts aus der DNA, ist in Abbildung 3.5 dargestellt. Sie läuft in vier Schritten ab. Im Einzelnen sind dieses:

1. Die DNA Doppelhelix wird durch ein Enzym teilweise aufgetrennt.
2. An einem der beiden Stränge der DNA wird durch das Enzym RNA-Polymerase ein neues RNA-Molekül gebildet. Hierbei werden die Regeln bzgl. der Komplementarität der Basen beachtet. Das kopierte RNA-Molekül, genannt Boten-RNA (mRNA), entspricht somit dem DNA-Molekül des komplementären DNA-Strangs.
3. Die Boten-RNA löst sich von der DNA.
4. Der DNA-Strang verbindet sich wieder und bildet die Doppelhelix.

Während der *Translation* wird auf der Basis der Boten-RNA ein Polypeptid synthetisiert. Hierbei wird die in der DNA als Nukleinsäure gespeicherte genetische Information in eine Kette von Aminosäuren umgewandelt. Der sogenannte *genetische Code* legt dabei fest welche Dreierkombination von Nukleotiden (welches *Triplett*) zu welcher Aminosäure umgewandelt wird.

In der Abbildung 3.6 ist die Translation der Boten-RNA in ein Protein dargestellt. Das gewundene Molekül, die *Transport-RNA* (tRNA), besitzt das Komplement zu einer Dreierkombination von Basen (das sogenannte *Anti-Codon*) auf der Boten-RNA. Zusätzlich ist die Transport-RNA mit einer Aminosäuren verbunden. Durch eine biochemische Reaktion wird die Aminosäure von der Transport-RNA getrennt und mit der bereits vorhandenen Aminosäurekette des entstehenden Polypeptids verbunden. Der gesamte Vorgang findet innerhalb der *Ribosomen* statt.

Nachdem alle Nukleotide der Boten-RNA in Aminosäuren übersetzt wurden steht das Polypeptid für die Zelle zur Verfügung und kann beispielsweise in seine dreidimensionale Struktur gefaltet werden.

3.3.2. Regulation der Genexpression

Viele der in den Genen kodierten Proteine werden nicht in allen Situationen benötigt. Das Bakterium *E. coli* etwa benötigt Enzyme zum Abbau von Laktose (Milchzucker) nur, wenn auch Laktose verfügbar ist und Pflanzen beispielsweise benötigen Gene zur Photosynthese

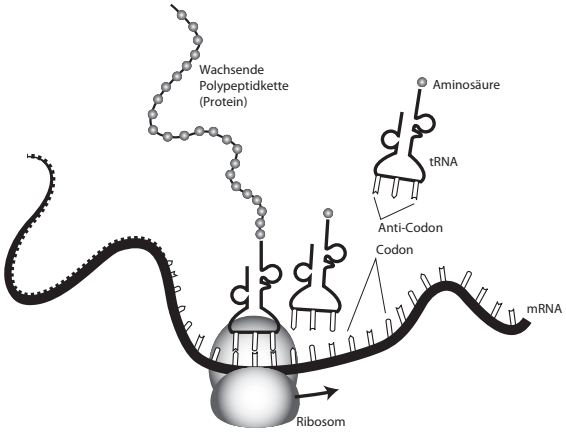


Abbildung 3.6.: Während der Translation wird zu einem mRNA-Strang ein Polypeptid gebildet. Hierzu verbindet sich jeweils ein tRNA-Molekül kurzzeitig mit der mRNA. Die Bindungsstelle ist durch das Triplett, dem *Codon* der mRNA bestimmt. Die tRNA hat auf der dem Anti-Codon gegenüberliegende Seite eine Aminosäure gebunden. Diese Aminosäure wird mit der wachsenden Polypeptidkette verbunden. Danach steht die tRNA wieder für weitere Translationsschritte zur Verfügung. Der gesamte Prozess der Translation findet in den Ribosomen statt. Bildquelle: [74]

nur, wenn Licht zur Verfügung steht. Die Expression der aktuell für eine Zelle notwendigen Gene wird deshalb durch unterschiedliche Mechanismen reguliert. Hierzu werden die einzelnen Schritte der Proteinsynthese durch verschiedene Prozesse verändert bzw. gesteuert. Der wichtigste Schritt in den regulierend eingegriffen wird ist dabei die Transkription eines Gens in die zugehörige mRNA.

Ein Gen hat hierzu einen so genannten *Promoter*. Hierbei handelt es sich um einen Bereich der DNA, der „vor“ dem eigentlichen Gen auf der DNA liegt. An diesen Promoter bindet das Protein RNA-Polymerase, welches die mRNA als Kopie der DNA erstellt. Reguliert wird die Expression eines Gens durch zusätzliche Regulatorproteine. Diese *Transkriptionsfaktoren* bezeichneten Proteine binden an die DNA und steuern auf diese Weise, ob die RNA-Polymerase an die DNA binden kann.

Es werden zwei Arten von Transkriptionsfaktoren unterschieden: aktivierende und reprimierende. Aktivierende Transkriptionsfaktoren (*Aktivatoren*) schalten das zugehörige Gen „ein“ und reprimierende Transkriptionsfaktoren (*Repressoren*) schalten es entsprechend „aus“. Hierzu binden die Transkriptionsfaktoren üblicherweise in der Nähe des Promoters an die DNA und verhindern im Fall reprimierender Transkriptionsfaktoren das Binden der RNA-Polymerase an die DNA. Diese kann folglich für das reprimierte Gen keine mRNA erstellen. Ein Aktivator hingegen ermöglicht die Bindung der RNA-Polymerase an die DNA, d.h. erst das Vorhandensein der Aktivatoren ermöglicht die Expression des betreffenden Gens.

Die Abbildung 3.7 zeigt ein Gen, den dazugehörigen Promoter und die Funktionsweise der RNA-Polymerase bei Vorhandensein bzw. Abwesenheit eines reprimierenden Transkriptionsfaktors.

Transkriptionsfaktoren selbst wiederum werden auf zwei Arten reguliert. Einerseits ist ein Transkriptionsfaktor selbst ein Protein. Das bedeutet, dass erst durch die Expression des entsprechenden Gens der Transkriptionsfaktor erzeugt wird und beispielsweise als Aktiva-

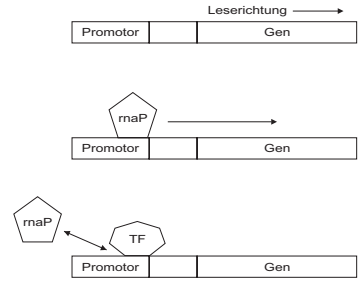


Abbildung 3.7.: (oben) Ein Gen und der dazugehörige Promoter. Das Gen wird in Richtung des Pfeils abgelesen. (mitte) Die RNA-Polymerase (rnaP) bindet in der Nähe des Promoters an die RNA und schreibt das Gen ab. (unten) Der (reprimierende) Transkriptionsfaktor bindet in der Nähe des Promoters. Die RNA-Polymerase kann folglich nicht an die RNA binden und das Gen kann nicht abgeschrieben werden.

tor wirken kann. Andererseits werden viele Transkriptionsfaktoren durch „Signalmoleküle“ aktiviert bzw. deaktiviert. Zu diesen Signalmolekülen gehören beispielsweise Hormone, aber auch Metaboliten, die in der Zelle auftreten. Laktose, wie oben schon genannt, ist unter gewissen Umständen für *E. coli* ein solches Signalmolekül. Wenn Laktose in der Zelle frei verfügbar ist, dann wird der reprimierende Transkriptionsfaktor *LacI* deaktiviert und die RNA-Polymerase kann die für die Verarbeitung der Laktose notwendigen Gene exprimieren.

3.3.3. Biochemische Reaktionen

Die in der Zelle ablaufenden Prozesse, beispielsweise die oben beschriebene Genexpression, benötigen Energie. Diese Energie gewinnt die Zelle aus dem Abbau von ihr zur Verfügung stehenden *Nährstoffmolekülen*. Beispielsweise ist Glukose ein Nährstoffmolekül für das Bakterium *E. coli*. Auf der anderen Seite erstellt die Zelle kontinuierlich neue Moleküle. Der oben beschriebene Prozess der Proteinsynthese beispielsweise nutzt die in der Zelle erstellte Energie um andere in der Zelle hergestellte Moleküle zu Polypeptidketten zu synthetisieren.

Der Abbau und die Synthese von Molekülen geschieht üblicherweise nicht in einem atomaren Schritt, sondern über mehrere Teilschritte. Jeder einzelne Teilschritt ist eine *biochemische Reaktion* und die Konkatenation mehrerer Reaktionen zu einer Kette bzw. einem (kleinen) Netzwerk wird *Stoffwechselweg* (engl. *Pathway*) genannt. Abbauende Stoffwechselwege werden *katabole Stoffwechselwege* und aufbauende werden *anabole Stoffwechselwege* genannt.

In der Abbildung 3.8 ist eine einzelne biochemische Reaktion, die durch das Enzym Hexokinase katalysiert wird, dargestellt. Diese Reaktion transformiert ein Molekül der α-D-Glukose unter Zuhilfenahme eines Moleküls Adenosintriphosphat (ATP) in ein Molekül α-D-Glukose-6-Phosphat. Dabei wird ein Molekül Adenosindiphosphat (ADP) freigesetzt.

Biochemische Reaktionen werden häufig durch *Enzyme* katalysiert. Das Enzym reduziert hierzu die als *Aktivierungsenergie* bezeichnete Schwelle, bei der die Reaktion ablaufen kann. Die Transformation der Eingangsstoffe (der *Substrate*) in die Ausgangsstoffe (die *Produkte*) wird folglich durch das Enzym vereinfacht bzw. beschleunigt. Nach Ablauf einer Transformation steht das Enzym üblicherweise wieder zur Verfügung und kann weitere Reaktionen katalysieren. Enzyme sind wiederum Proteine die durch den oben beschriebenen Prozess der Proteinsynthese erstellt werden.

Biochemische Reaktionen benötigen teilweise mehr als ein Molekül eines Stoffes. Übli-

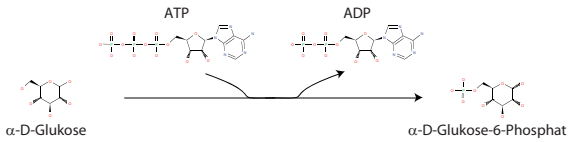


Abbildung 3.8.: Die durch das Enzyme Hexokinase katalysierte Reaktion, dargestellt als Hypergraph. In dieser Reaktion wird α-D-Glukose unter Verwendung von Adenosintriphosphat (ATP) zu α-D-Glukose-6-Phosphat umgewandelt, hierbei wird Adenosindiphosphat (ADP) frei.

cherweise wird deshalb in der Beschreibung einer Reaktion angegeben, wie viele Moleküle jeweils notwendig sind. Für die Aufspaltung von Pyrophosphat in Orthophosphat gilt beispielsweise: $H_4P_2O_7 + H_2O \Leftrightarrow 2\ H_3PO_4$. Durch die beschriebene Reaktion wird also ein Molekül mit zwei Phosphoratomen (Pyrophosphat, $H_4P_2O_7$) und ein Molekül Wasser (H_2O) in zwei Moleküle mit je einem Phosphoratom (Orthophosphat, H_3PO_4) umgewandelt. Die Information über die mengenmäßige Zusammensetzung der Reaktion wird als *Stöchiometrie* bezeichnet.

Reaktionen, die nur in eine Richtung ablaufen können, werden als *irreversible Reaktionen* und Reaktionen, die in beide Richtungen ablaufen können, werden als *reversible Reaktionen* bezeichnet.

3.3.4. Interaktionen zwischen Proteinen

Proteine wirken innerhalb einer Zelle nicht einzeln, sondern treten üblicherweise in Wechselwirkung mit anderen Proteinen oder anderen Molekülen. Zu diesen *Wechselwirkungen*, auch *Protein-Interaktionen* genannt, gehört beispielsweise die Funktionsweise der oben beschriebenen Transkriptionsfaktoren. Auch die Phosphorylierung, eine biochemische Modifikation, eines Proteins durch ein anderes Protein, ist eine Protein-Wechselwirkung. Eine dritte Variante der Protein-Wechselwirkungen ist die Bildung von Proteinkomplexen. Hierbei handelt es sich um eine Verbindung, die mehrere, auch unterschiedliche, Proteine eingehen, um gemeinsam ein Funktion zu erfüllen bzw. ein größeres Molekül zu bilden. Hämoglobin, das für den Sauerstofftransport im Blut notwendige Protein, ist beispielsweise ein Proteinkomplex bestehend aus vier Proteinen, zwei α- und zwei β-Hämoglobin Proteinen.

3.4. Netze des Lebens

Innerhalb der Zelle laufen die beschriebenen Prozesse (Regulation der Genexpression, biochemische Reaktionen und Protein-Interaktionen) nicht unabhängig voneinander ab. Beispielsweise können die Produkte einer biochemischen Reaktion als Substrate einer anderen („nachfolgenden“) Reaktion verwendet werden.

Die durch die Zusammenschaltung der drei oben genannten Prozesse entstehenden Netzwerke werden im folgenden Abschnitt beschrieben.

3.4.1. Genregulationsnetze

Ein Großteil der Regulation der Genexpression findet auf der Ebene der Transkription (siehe auch Abschnitt 3.3.2) statt. Hierbei übt ein Transkriptionsfaktor einen regulatorischen

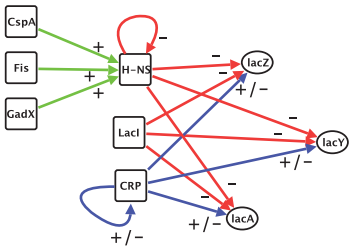


Abbildung 3.9.: Ausschnitt aus der transkriptionellen Regulation der drei in *E. coli* für die Laktose-Verarbeitung notwendigen Proteine. Rechtecke repräsentieren Gene, die Transkriptionsfaktoren kodieren, Ellipsen repräsentieren Gene, die andere Proteine (Zielgene) kodieren, aktivierende Einflüsse sind als grüne Kanten mit dem Kantenlabel +, reprimierende Einflüsse als rote Kanten mit dem Kantenlabel − und duale Einflüsse als blaue Kanten mit dem Kantenlabel +/− dargestellt. Datenquelle: RegulonDB [147], Visualisierung: VANTED [81]

Einfluss auf die Expression eines oder mehrerer Zielgene aus. Dieser Einfluss kann entweder positiv sein, d.h. die Expression des Zielgens verstärken, oder negativ sein, d.h. die Expression verringern.

Modelliert wird dieser Einfluss eines Transkriptionsfaktors auf ein Zielgen als gerichteter Graph mit zwei Knoten und einer Kante. Der Startknoten der Kante repräsentiert dabei das den Transkriptionsfaktor kodierende Gen und der Zielknoten das regulierte Gen. Transkriptionsfaktoren selbst sind Genprodukte und unterliegen somit dem regulatorischen Einflüssen durch andere Transkriptionsfaktoren. Hierdurch wird unmittelbar ein gerichteter, gelabelter Graph, das *Genregulationsnetz*, aufgespannt.

In der Abbildung 3.9 ist ein Ausschnitt aus dem Genregulationsnetz des Bakteriums *E. coli* dargestellt. Sichtbar ist der Einfluss der Transkriptionsfaktoren *H-NS*, *LacI* und *CRP* auf die drei für die Laktose-Verarbeitung notwendigen und durch die Gene *lacZ*, *lacY* und *lacA* kodierten Proteine. Das den Transkriptionsfaktor *H-NS* kodierende Gen wird wiederum von drei Transkriptionsfaktoren (*CspA*, *Fis* und *GadX*) reguliert.

Genregulationsnetze werden auf der Basis von beschriebenen Eigenschaften von Transkriptionsfaktoren erstellt. Hierzu existieren verschiedene Datenbanken, die die entsprechenden Informationen, häufig unmittelbar mit dem entsprechenden Verweis auf die dazugehörige Literaturstelle, bereitstellen. Beispiele für Datenbanken, die Informationen über die transkriptionelle Regulation bereitstellen sind: TransFac für Eukaryoten [88, 122], RegulonDB für *E. coli* [147] und AtRegNet für *Arabidopsis thaliana* [132].

In der Abbildung 3.10 ist ein Genregulationsnetz für das Bakterium *E. coli* dargestellt. Dieses Netz wird im Kapitel 5 verwendet.

3.4.2. Metabolische Reaktionsnetze

Biochemische Reaktionen wandeln Substrate, häufig unter dem Einfluss eines katalysierenden Enzyms, in Produkte um (siehe auch Abschnitt 3.3.3). Diese Umwandlung wird in der Literatur üblicherweise als *Hypergraph*, d.h. als Graph, in dem Kanten mehr als einen Start- und mehr als einen Endknoten haben dürfen, dargestellt. Für die üblichen Analyseverfahren, beispielsweise für die Zentralitätsanalyse, eignen sich Hypergraphen jedoch nur bedingt. Folglich werden die Reaktionen als bipartite Graphen, bei denen die Metaboliten die eine und die Reaktionen die andere Knotenmenge bilden, modelliert. Verbunden sind jeweils die

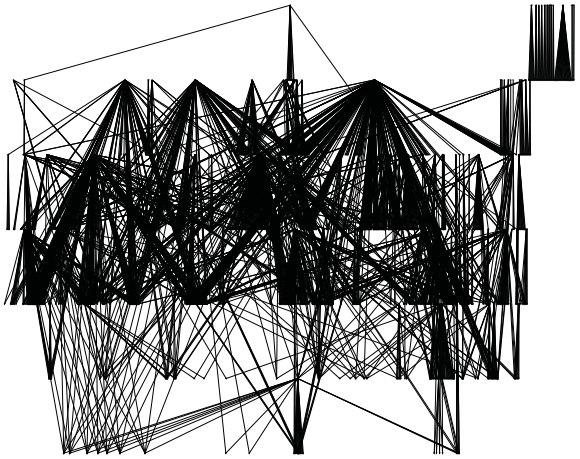


Abbildung 3.10.: Ein Genregulationsnetz für *E. coli* bestehend aus 1250 Knoten (Genen, die Transkriptionsfaktoren bzw. andere Proteine kodieren) und 2515 Kanten. Deutlich erkennbar ist ein hierarchischer Aufbau des Netzes: Einige Transkriptionsfaktoren regulieren direkt und indirekt viele andere Gene. Datenquelle: RegulonDB (Version 5.0) [147], Visualisierung: Graphviz (Hierarchisches Layout) [59, 60]

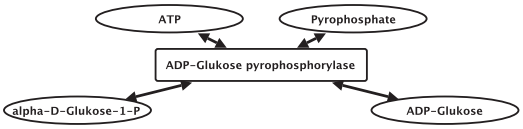


Abbildung 3.11.: Die durch das Enzym ADP-Glukose Pyrophosphorylase katalysierte Reaktion modelliert als bipartiter Graph. Metaboliten sind als Ellipsen und die Reaktion ist als Rechteck dargestellt. Da die Reaktionsrichtung nicht bekannt ist, ist die Reaktion als reversible Reaktion, d.h. als ungerichteter Graph, dargestellt. In dieser Darstellungart ist allerdings nicht erkennbar welche Metaboliten Substrate und welche Metaboliten Produkte der Reaktion sind. Visualisierung: VANTED [81]

Substrat-Metaboliten über Kanten mit der Reaktion und die Reaktion ist mit den Produkt-Metaboliten ebenfalls durch entsprechende Kanten verbunden, siehe Abbildung 3.11.

Die Richtung, in der eine biochemische Reaktion abläuft, wird häufig erst durch die Umgebung innerhalb der Zelle bestimmt. Dieser Umgebungszustand ist bei der Modellierung von Reaktionen als Graph nicht notwendigerweise bekannt. Üblicherweise wird deshalb für eine Reaktion angenommen, dass diese reversibel ist. Um dennoch die Trennung in Substrat- und Produkt-Metaboliten zu ermöglichen, werden die reversiblen Reaktionen dupliziert. Hierzu wird eine Reaktion, die die Substrate in die Produkte umwandelt, und eine zweite Reaktion, die die Produkte in die Substrate, umwandelt modelliert. Für die durch das Enzym ADP-Glukose Pyrophosphorylase katalysierte Reaktion ist dieses Art der Modellierung in Abbildung 3.12 dargestellt.

Die Stöchiometrie einer Reaktion wird bei der Modellierung über Kantengewichte abgebildet. Bei der in Abschnitt 3.3.3 beschriebenen Umwandlung von Pyrophosphat in Orthophosphat ist somit das Kantengewicht der Kanten zwischen den beiden Substraten und der Reaktion jeweils 1 und zwischen der Reaktion und dem Produkt 2.

Produkt-Metaboliten einer Reaktion können als Substrat-Metaboliten einer anderen Re-

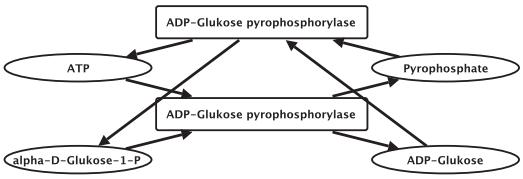


Abbildung 3.12.: Die durch das Enzym ADP-Glukose Pyrophosphorylase katalysierte Reaktion modelliert als bipartiter, gerichteter Graph in der die reversible Reaktion dupliziert ist. Beide Richtungen sind in dieser Darstellung korrekt modelliert, d.h. es ist klar erkennbar welche Metaboliten die Substrate und welche Metaboliten die Produkte sind. Visualisierung: VANTED [81]

aktion fungieren. Folglich ergibt sich durch die Verbindung der Metaboliten unmittelbar ein Graph. In der Abbildung 3.13 ist die Glykolyse für das Bakterium *E. coli*, wie sie in der Datenbank EcoCyc [90] beschrieben ist, als bipartiter Graph dargestellt.

Metaboliten, die in einer biochemischen Reaktionen nur eine „Nebenrolle“ spielen, werden üblicherweise als *Kofaktoren* bezeichnet. Häufig werden bei der Darstellung von Reaktionsnetzen diese Kofaktoren mehrfach aufgeführt, da dieses zu einer übersichtlicheren Zeichnung führt. In der Abbildung 3.13 sind beispielsweise die Kofaktoren ATP, ADP, Orthophosphat und H_2O jeweils mehrfach dargestellt.

Durch die Projektion des bipartiten Graphen auf eine der beiden Knotenmengen entsteht der zugehörige *Metabolit-* bzw. *Reaktionsgraph* (oder auch *Enzymgraph*) zum gegebenen Reaktionsnetz (siehe auch Def. 2.17). Aus diesem ist ablesbar, welches Metabolit in welches andere Metabolit umgewandelt werden kann bzw. welche Reaktion als Vorgänger bzw. Nachfolger einer anderen Reaktion auftritt. In der Abbildung 3.14 ist der Metabolit- und der Reaktionsgraph zur Glykolyse in *E. coli* (Abbildung 3.13) dargestellt.

Beschreibungen biochemischer Reaktionen sind in Datenbanken, beispielsweise KEGG LIGAND [65], hinterlegt und in der Literatur [20, 124] zu finden. Die Reaktionsbeschreibungen sind dort allerdings üblicherweise nicht organismenspezifisch, d.h. es ist nicht unbedingt ersichtlich, ob die beschriebene Reaktion auch tatsächlich im zu untersuchenden Organismus abläuft. Hierzu existieren Datenbanken, die entweder handverlesene Reaktionsbeschreibungen für einzelne Organismen enthalten oder eine automatische Ableitung der vorhandenen Reaktionen anhand von Informationen über die Funktionsweise bekannter Gene vornehmen. Die zur Zeit am IPK in Gatersleben entstehende Datenbank MetaCrop [66, 169] gehört zu den Datenbanken, die handverlesene Reaktionsbeschreibungen enthalten. In diese Datenbank werden detaillierte Reaktionsbeschreibungen für verschiedene Nutzpflanzen, beispielsweise *Hordeum vulgare* (Gerste), eingepflegt. Die Datenbank KEGG Pathway [84] gehört hingegen zu den Datenbank, bei denen für die Bestimmung des Vorhandenseins einer Reaktion die Annotation, d.h. die Beschreibung der Funktion der Gene des betreffenden Organismus, genutzt wird. Wenn zu einer gesuchten Reaktion ein Gen, dessen Genprodukt als Enzym die betrachtete Reaktion katalysiert gefunden wird, dann wird angenommen, dass die betreffende Reaktion im Organismus ablaufen kann.

Die Abbildung 3.15 zeigt einen Teil des biochemischen Reaktionsnetzes von *E. coli*. Dieses Netz wird in Kapitel 6 analysiert.

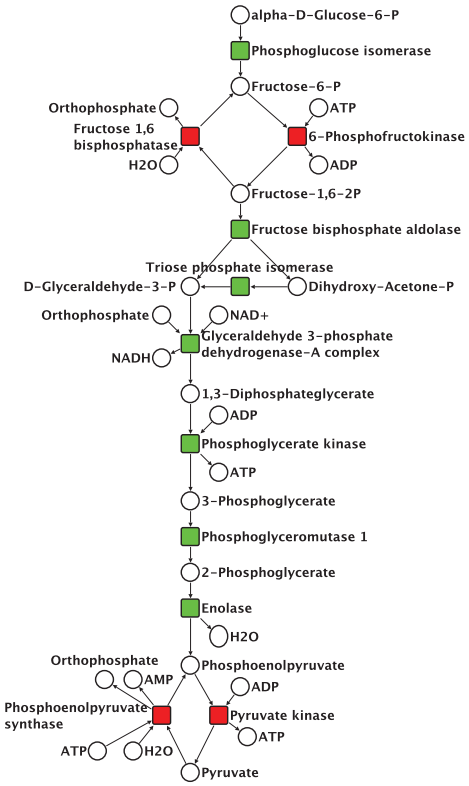


Abbildung 3.13.: Der Stoffwechselweg *Glykolyse*, d.h. der Abbau der Glukose zur Gewinnung von Energie, für *E. coli*. Metaboliten sind als Kreis und Reaktionen als Rechteck dargestellt. Die in der Datenbank als irreversibel angegebenen Reaktionen sind rot und die als reversibel angegebenen Reaktionen grün gekennzeichnet. Zu den reversiblen, in grün dargestellten, Reaktionen sind die entsprechenden dualen Reaktionen nicht dargestellt. Einige Kofaktoren sind mehrfach als Knoten im Netzwerk dargestellt und die entsprechenden Kanten führen nur zum nächstgelegenen Knoten. Diese Darstellungsvariante führt zu einer klareren Visualisierung. Datenquelle: EcoCyc [90], Visualisierung: VANTED [81]

3.4.3. Protein-Interaktionsnetze

Wechselwirkungen zwischen Proteinen und anderen Molekülen mit Proteinen können innerhalb eines Organismus auf vielfältige Weise auftreten. Einige Beispiele sind im Abschnitt 3.3.4 aufgeführt. Da Proteine in mehr als eine Interaktion mit anderen Proteinen eintreten können, lässt sich auch die Menge der Protein-Interaktionen als Graph, dem *Protein-Interaktionsnetz*, darstellen.

Der sehr allgemeine Begriff der Wechselwirkung zwischen zwei Objekten lässt eine Aussage nach der Art oder der Reihenfolge der Interaktion nicht zu. Protein-Interaktionen werden aus diesem Grund als ungerichtete Graphen modelliert.

In der Abbildung 3.16 sind die Interaktionen der beiden Proteine α-Hämoglobin und β-Hämoglobin dargestellt. Da beide einen Proteinkomplex bilden, sind diese über eine Kante miteinander verbunden. In der verwendeten Datenbank IntAct [89] sind zusätzlich noch 6 Interaktionen für α- und 2 Interaktionen für β-Hämoglobin verzeichnet.

Informationen über Interaktionen von Proteinen sind in einer ganzen Reihe von Datenbanken hinterlegt. Die Internetseite Pathguide [16] führte im Mai 2007 insgesamt 88

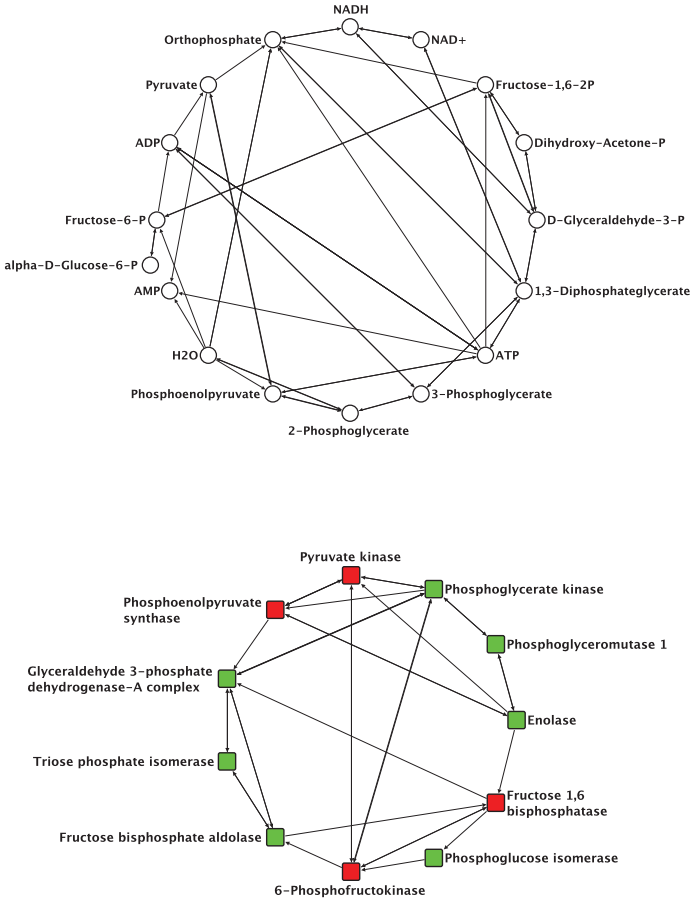


Abbildung 3.14.: Metabolit- (oben) und Reaktionsgraph (unten) der Glykolyse (Abbildung 3.13) für *E. coli*. Datenquelle: EcoCyc [90], Visualisierung: VANTED [81]

unterschiedliche Datenbanken in der Kategorie Protein-Interaktionen auf. Nicht alle dieser Datenbanken enthalten allerdings ausschließlich Interaktionen zwischen zwei Proteinen. Manche, beispielsweise BIND [15], enthalten auch Interaktionen zwischen Proteinen und anderen Molekülen oder Interaktionen, die aufgrund von Ähnlichkeit in der Protein- oder Nukleotidsequenz vorhergesagt wurden [179].

In der Abbildung 3.17 ist ein Teil des Protein-Interaktionsnetzes für *E. coli*, extrahiert aus der Datenbank DIP [148], dargestellt.

3.5. Verwendete Literatur

Die einzelnen Bausteine und Prozesse der Molekularbiologie sind in den Büchern von Alberts und Koautoren [4] und Clark und Koautoren [29] sehr anschaulich und detailliert beschrieben. Molekularbiologische Netzwerke werden bei Kanehisa [83], Klipp und Koautoren [91] und Palsson [133] erörtert.

Eine große Anzahl von Überblicksartikeln sind zu verschiedenen Bereichen der Analyse molekularbiologischer (und anderer biologischer) Netzwerke erschienen. Etwas allgemeineren Fokus haben dabei die Arbeiten von Aittokallio & Schwikowski [2], Albert [3], Barabási & Oltvai [17], Mason & Verwoerd [121] und Proulx, Promislow & Phillips [136]. Auf die Metho-

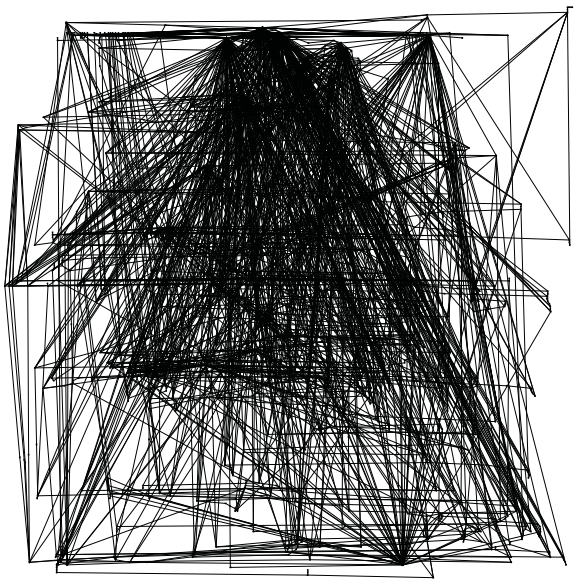


Abbildung 3.15.: Ausschnitt aus dem biochemischen Reaktionsnetz von *E. coli* modelliert als bipartiter Graph. Das Netzwerk besteht aus 538 Metaboliten, 626 Reaktionen und 2419 Kanten. Datenquelle: [41], Visualisierung: Graphviz (Hierarchisches Layout) [59, 60]

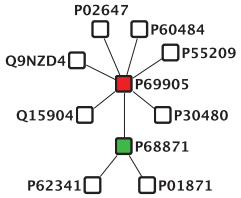


Abbildung 3.16.: Protein-Interaktionen von α-Hämoglobin (P69905) und β-Hämoglobin (P68871) für *Homo sapiens*. Die Benennungen der Knoten entsprechen den Identifikatoren der entsprechenden Einträge in der Datenbank UniProt [164]. Datenquelle: IntAct [89], Visualisierung: VANTED [81]

den zur experimentellen und computerbasierten Herleitung von Netzwerken konzentrieren sich Xia und Koautoren [176]. Einen Schwerpunkt auf Signalübertragungsnetze und Modularität in diesen Netzwerken legen Qi & Ge [138]. Almaas [6] und Zhu und Koautoren [181] wiederum fokussieren auf die strukturellen Eigenschaften der Netzwerke.

3.6. Zusammenfassung

In diesem Kapitel wurden drei Arten von molekularbiologischen Netzwerken und die diesen zugrunde liegenden Prozesse und Bausteine beschrieben. Alle drei Netzwerktypen wurden bereits mittels Zentralitäten analysiert und für die beiden Netzwerktypen Genregulationsnetze und metabolische Reaktionsnetze wird jeweils in einem der späteren Kapitel eine neue Zentralität vorgestellt.

Die Tabelle 3.1 fasst die für eine Analyse mittels Zentralitäten wesentlichen Grapheigenschaften der beschriebenen Netzwerktypen zusammen. Da im Folgenden Metabolit- und Reaktionsgraphen noch mehrfach verwendet werden, sind diese in dieser Tabelle separat mit

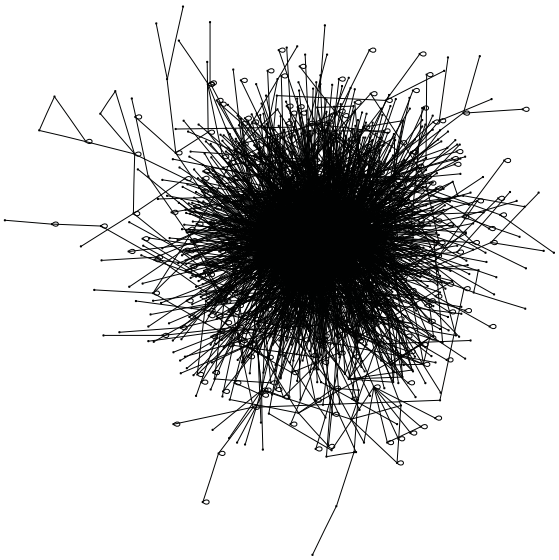


Abbildung 3.17.: Die größte Zusammenhangskomponente des Protein-Interaktionsnetzes des Bakteriums *E. coli*. Das dargestellte Netz besteht aus 1413 Knoten und 6535 Kanten. Datenquelle: Database of Interacting Proteins (DIP, Stand: 7. Januar 2007) [89], Visualisierung: Graphviz (Kräftebasiertes Layout) [60, 82]

aufgeführt.

Bezeichnung	Graphtyp	Schleifen	Zusammenhang	Kantengewicht
Genregulationsnetz	gerichtet	möglich	schwach	$+1/-1$[a]
Protein-Interaktionsnetz	ungerichtet	möglich	zshg.	Noch keine[b]
Metab. Reaktionsnetz	gerichtet, bipartit	verboten	schwach	Stöchiometrie[c]
Metabolitgraph	gerichtet, unipartit	möglich	schwach	keine[d,e]
Reaktionsgraph	gerichtet, unipartit	möglich	schwach	keine[d]

[a] Aktivierungen bzw. Repressionen werden üblicherweise als $+1$ bzw. -1 modelliert.
[b] Zur Zeit sind keine Datenbanken verfügbar, die „Interaktionsstärken“ o.ä. bereitstellen.
[c] Die Anzahl der an einer Reaktion beteiligten Metaboliten kann als Kantengewicht repräsentiert werden.
[d] Bisher wurden keine Kantengewichte bei diesem Netzwerktyp verwendet.
[e] Für diesen Netzwerktyp wird im Kapitel 6 eine Zentralität, die auch Kantengewichte verwendet, eingeführt.

Tabelle 3.1.: Für die molekularbiologischen Netzwerke sind die für die Zentralitätsanalyse wichtigsten Grapheigenschaften in dieser Tabelle dargestellt. Die zweite Spalte beschriebt die generelle Struktur des Graphen. Ob dieser Netzwerktyp Schleifen haben kann, ist in der dritten Spalte angegeben. Der üblicherweise bei einer Zentralitätsanalyse verwendet Zusammenhangsbegriff ist in der vierten Spalte notiert. Die Interpretation möglicher Kantengewichte ist in der letzten Spalte aufgeführt.

4. Motivation für die Vorstellung neuer Zentralitätsmaße zur Analyse molekularbiologischer Netzwerke

In zahlreichen Publikationen wurden bereits Zentralitäten zur Analyse molekularbiologischer Netzwerke eingesetzt. In diesem Kapitel werden deshalb die in diesen Veröffentlichungen beschriebenen Resultate überblicksartig zusammengefasst. Hierbei werden zuerst Genregulationsnetze, danach metabolische Reaktionsnetze und abschließend Protein-Interaktionsnetze betrachtet. Darauf aufbauend wird motiviert, warum zur Analyse molekularbiologischer Netzwerke neue bzw. angepasste Zentralitätsmaße notwendig sind.

4.1. Publizierte Zentralitätsanalysen von Genregulationsnetzen

Für die Analyse von Genregulationsnetzen wurden bisher die folgenden Zentralitätsmaße eingesetzt: *SP-Betweenness*, *Degree*, *Eccentricity*, *Closeness*, *CF-Betweenness* und *Eigenvector*-Zentralität.

Ein nicht öffentlich verfügbares Genregulationsnetz für Säugetiere (engl. *mammalians*) wurde von Potapov und Koautoren analysiert [135]. Dieses Netzwerk wurde auf der Basis der Informationen aus den Datenbanken TRANSFAC [122] und TRANSPATH [100] erstellt. Zur Bestimmung wichtiger Transkriptionsfaktoren wurde die *Shortest-Path Betweenness* eingesetzt. Transkriptionsfaktoren mit hohem Betweenness-Wert in diesem Netz, beispielsweise *p53*, *c-fos*, *c-jun*, *SRF* und *c-myc*, sind alle als Regulatoren in der Zellvermehrung bekannt und es ist bekannt, dass es sich bei den zugehörigen Genen um Tumorsuppressorgene[1] bzw. Protoonkogene[2] handelt. Beide Arten von Genen spielen in der Krebsforschung eine wichtige Rolle.

Insgesamt fünf Zentralitäten (*Degree*, *Eccentricity*, *Closeness*, *Random-Walk Betweenness* und *Eigenvector*-Zentralität) wurden von Koschützki und Schreiber auf ein (ungerichtetes) Genregulationsnetz von *E. coli* angewendet [97]. Es wurde beobachtet, dass die einzelnen Zentralitätsmaße miteinander korrelieren und das insbesondere die Korrelation zwischen den Zentralitäten *Eigenvector* und *Closeness* stark positiv ist ($r = 0,9552$, Pearsonscher Maßkorrelationskoeffizient [93]).

In der Tabelle 4.1 sind die in diesem Abschnitt diskutierten Publikationen zusammengefasst dargestellt.

[1] Ein Gen, das die Wahrscheinlichkeit verringert, das sich eine Zelle in eine Tumorzelle verwandelt.
[2] Ein Gen, das durch Mutationen zu einem Krebs-Gen mutieren kann.

Zentralität	*E. coli*	Säugetiere
Degree	[97]	—
Eccentricity	[97]	—
Closeness	[97]	—
Centroid-Value	—	—
Load Point	—	—
CF-Closeness	—	—
Subgraph	—	—
Bipartivity	—	—
Transitivity	—	—
SP-Betweenness	—	[135]
CF-Betweenness	[97]	—
Eigenvector	[97]	—
Entropy	—	—
Damage	—	—
Avalanche	—	—

Tabelle 4.1.: Aufstellung existierender Publikationen, in denen Genregulationsnetze mittels Zentralitäten analysiert wurden. In den Zeilen sind die in diesem Dokument beschriebenen Zentralitäten aufgeführt und in den Spalten die Namen der Organismen, die in den benannten Publikationen analysiert wurden.

4.2. Publizierte Zentralitätsanalysen von metabolischen Reaktionsnetzen

Für die Analyse von metabolischen Reaktionsnetzen wurden bisher die folgenden Zentralitätsmaße eingesetzt: *Degree*, *Eccentricity*, *Closeness*, *Centroid Value*, *Load Point*, *SP-Betweenness* und *Avalanche*.

Im Folgenden (siehe auch Tabelle 4.2) werden die wesentlichen Resultate der bereits veröffentlichten Analysen vorgestellt. Dabei wird unterschieden, ob eine Reihenfolge von Metaboliten oder eine Reihenfolge von Reaktionen bzw. Enzymen erstellt wurde. Zusätzlich wurden von einigen Autoren Zentralitäten zur Beschreibung bzw. zum Vergleich der Struktur von Netzwerken für unterschiedliche Organismen verwendet.

4.2.1. Reihenfolgen von Metaboliten

Die *Degree*-Zentralität wurde von Jeong und Koautoren im Rahmen einer Analyse der Netzwerktopologie der metabolischen Reaktionsnetze von insgesamt 43 Organismen eingesetzt [79]. Die verwendeten Daten für diese Analyse wurden dazu aus der Datenbank WIT [129] extrahiert. Pro Organismus wurde ein bipartiter Graph auf der Basis der Reaktionsgleichungen erstellt, und verschiedene topologische Eigenschaften der Graphen wurden untersucht. Eines der Resultate ist die Bestimmung der Metaboliten, die an vielen Reaktionen, entweder als Substrate (*Out-Degree*) oder als Produkte (*In-Degree*) beteiligt sind. Hierbei wurde beobachtet, dass in fast allen betrachteten Organismen die Reihenfolge der Metaboliten gemäß des *In-* bzw. *Out-Degree*s nahezu übereinstimmt. Aufgrund der gewählten Modellierung des Netzes ist dieses Resultat aber nicht überraschend: die Kofaktoren, beispielsweise ATP, ADP, NAD, NADH und auch H_20, stehen bei fast allen Organismen auf den vorderen Plätzen. Dieses lässt sich anhand der Biochemie einfach erklären: Kofaktoren treten zwar in unterschiedlichen Reaktionen, hierbei aber jeweils mit derselben biochemi-

schen Funktion, auf. Infolgedessen werden die Kofaktoren gemäß der *Degree*-Zentralität als wichtig identifiziert.

Wagner & Fell haben ein metabolisches Reaktionsnetz von *E. coli* untersucht [50, 49, 167]. Neben dem *Degree* wurde die *Closeness*-Zentralität (hier *Importance Number* genannt) zur Erstellung einer Reihenfolge der Metaboliten genutzt. Berechnungsgrundlage ist der ungerichtete Metabolitgraph (siehe Abschnitt 3.4.2) eines Ausschnitts des Metabolismus von *E. coli*. Bei der Erstellung des Graphen wurde eine Reihe von Metaboliten nicht berücksichtigt. Die Liste der entfernten Metaboliten umfasst die hauptsächlich als Kofaktoren auftretenden Metaboliten ATP, ADP, NAD, NADH, NADP, NADPH, CO_2, NH_3, SO_4, Thioredoxin, organisches Phosphat und Pyrophosphat. Die für die Zentralitäten *Degree* und *Closeness* publizierten Reihenfolgen der Metaboliten sind sich ähnlich. Dominiert wird die angegebene Liste der Top 13 Metaboliten jeweils von Metaboliten des Zitronensäurezyklus, insbesondere wenn Aminosäuren die durch Transaminierung[3] aus den Zitronensäurezyklus-Metaboliten erstellt werden, mitgezählt werden.

Wuchty & Stadler haben die beiden von Jeong und Koautoren bzw. Fell & Wagner (ohne die Bereinigung um die Kofaktoren) verwendeten Reaktionsnetze für *E. coli* erneut analysiert [174]. Zum Einsatz kamen die drei Zentralitäten *Eccentricity*, *Closeness* und *Centroid Value*. Die Anwendung dieser Zentralitäten auf das als bipartiter Graph modellierte Reaktionsnetz (Jeong *et al.*) und auf den Metabolitgraphen (Fell & Wagner) lieferte identische Resultate: unter den Metaboliten mit den höchsten Zentralitätswerten befinden sich viele bekannte Kofaktoren wie beispielsweise ATP, ADP und AMP, NAD, NADH, NADP und NADPH, Ortho- und Pyrophosphat und auch CO_2. Dieses Resultat ist nicht überraschend: beide analysierten Graphen haben eine Gradverteilung, die das jeweilige Netz als *scale-free network*[4] (deutsch „skalenfreies Netz“) klassifiziert. Das bedeutet, dass eine Reihe von Knoten, nämlich genau die Kofaktoren, einen herausragend hohen Knotengrad haben. Von diesen Knoten existieren folglich viele kurze Wege zu den anderen Knoten.

Auf der Basis der LIGAND-Datenbank, einer Teildatenbank der *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*) [65, 84], haben Ma & Zeng die Reaktionsnetze von 80 Organismen erstellt [110, 111]. Hierzu wurden jeweils die einzelnen Reaktionen betrachtet und (a) entschieden, ob die Reaktion gerichtet oder ungerichtet ist, (b) (falls möglich) die Reaktionsrichtung festgelegt und (c) eventuell in der Reaktion auftretende Kofaktoren identifiziert und entfernt. Aus diesen Daten wurde für jeden Organismus ein Metabolitgraph erstellt. Die so erstellten Graphen unterscheiden sich von den in vorigen Publikationen verwendeten deutlich: das Löschen der Kofaktoren aus den Reaktionsgleichungen und das zusätzliche Betrachten der Reaktionsrichtung führt zu einer veränderten Netzstruktur. Sichtbar wird diese bereits bei der Betrachtung der mittleren Pfadlänge[5] (*Average Path Length*). Für *E. coli* liegt dieser Wert für die Netze von Jeong *et al.* bzw. Fell & Wagner bei 3,2 bzw. 3,8. Ma & Zeng geben diesen Wert hingegen mit 8,2 an. Ma & Zeng geben auch eine Liste mit den Top-10 Metaboliten gemäß der *In-* und *Out-Closeness*-Zentralität für *E. coli* an. Acht der 10 Metaboliten auf dieser Liste gehören zu den Stoffwechselwegen Glykolyse und Zitronensäurezyklus. Das Metabolit mit dem höchsten Zentralitätswert ist Pyruvat, das Metabolit am Übergang von der Glykolyse in den Zitronensäurezyklus.

Arita analysierte ebenfalls die Netzwerktopologie eines Metabolitgraphen für *E. coli* [14]. Im Unterschied zu den vorigen Analysen wurde das Netzwerk auf der Basis von Informatio-

[3] Eine Verschiebung einer Aminogruppe (NH_2) von einer α-Aminosäure an eine α-Ketosäure.

[4] In einem *scale-free network* haben einige Knoten (*hubs*) sehr viele Kanten und die große Mehrzahl der Knoten haben nur einige wenige Kanten.

[5] Definiert als der Mittelwert über alle Distanzen (siehe Definition 2.8) innerhalb des Graphen.

nen über die Verwendung der einzelnen Kohlenstoffatome innerhalb der Reaktionen erstellt. Je zwei Metaboliten wurden miteinander verbunden, wenn eine Reaktion existiert, in der mindestens ein Kohlenstoffatom von einem Metabolit an ein anderes Metabolit übertragen wird. Durch diese Modellierung wird die Bedeutung der Kofaktoren erheblich reduziert, da diese häufig keine Kohlenstoffatome mit den Reaktionspartnern austauschen. Die bei Arita angegebene Liste der Top 10 Metaboliten anhand des *Degrees* enthält deshalb hauptsächlich Metaboliten, die Kohlenstoffatome an viele andere Metaboliten abgeben können bzw. von diesen erhalten. Konsequenterweise wird die Liste von CO_2, Pyruvat und Acetyl-Coenzym A angeführt.

Mahadevan & Palsson betrachteten die Metabolitgraphen der metabolischen Reaktionsnetze von drei Organismen: *E. coli*, *S. cerevisiae* und *G. sulfurreducens* [114]. Sie berechneten den *Degree* der Metaboliten und verglichen diesen mit der Wahrscheinlichkeit, dass die daran verbundenen Kanten (Reaktionen) lebensnotwendig sind. Eine Kante gilt hierbei als lebensnotwendig, wenn der entsprechende Organismus nach einer Deaktivierung des zum katalysierenden Enzym gehörenden Gens nicht mehr lebensfähig ist. Es zeigte sich bei dieser Analyse, dass allein aus dem *Degree* des Metaboliten nicht auf die Lebensnotwendigkeit der verbundenen Reaktionen geschlossen werden kann. In jedem der betrachteten Organismen gibt es Metaboliten mit einem geringen *Degree*, bei der alle Reaktionen in denen das Metabolit als Substrat bzw. Produkt vorkommt lebensnotwendig sind. Hingegen gilt, dass aus dem *Degree* des Metaboliten auf die Lebensnotwendigkeit des Metaboliten geschlossen werden kann: wenn ein Metabolit mit einem hohen *Degree* aus dem Netz entfernt wird, dann werden hierdurch auch die entsprechenden Reaktionen (Kanten) entfernt. Die Wahrscheinlichkeit, dass hierbei eine lebensnotwendige Reaktion entfernt wird steigt dabei mit dem *Degree* des betrachteten Metaboliten.

Von Rahman & Schomburg wurde die Zentralität *Load Point* für die Analyse metabolischer Reaktionsnetze vorgestellt. Mit dieser Zentralität kann sowohl für die Metaboliten als auch für die Reaktionen eines metabolischen Reaktionsnetzes eine Reihenfolge ermittelt werden. Hierzu wird jeweils der zugehörige unipartite Graph zum als bipartiter Graph modellierten Reaktionsnetz erstellt und die in Definition 2.24 beschriebene Zentralität berechnet. Angewendet auf einen Metabolitgraphen lässt sich somit eine Reihenfolge der Metaboliten ermitteln. Rahman & Schomburg haben die Zentralität zur Bewertung der Metaboliten je eines Reaktionsnetzes für die Organismen *Bacillus subtilis 168* und *Bacillus antracis Sterne* verwendet [140]. Die Liste der Top 10 Metaboliten der beiden Organismen unterscheidet sich dabei nur in der Sortierung und nicht in der grundsätzlichen Zusammensetzung. Nur eines der Metaboliten (*Cystathionine*), das in der Reihenfolge für *B. subtilis 168* an Position 5 steht, kommt in der entsprechenden Reihenfolge des anderen Bakteriums nicht vor. Eine detailliertere Diskussion der ermittelten Reihenfolge der Metaboliten erfolgt bei Rahman & Schomburg nicht.

4.2.2. Reihenfolgen von Reaktionen bzw. Enzymen

Liu & Koautoren haben untersucht, ob das phylogenetische Profil von Enzymen mit unterschiedlichen Zentralitätswerten derselben Enzyme korrelieren [106]. Das phylogenetische Profil eines Enzyms ist hierbei definiert als die Häufigkeit des Vorkommens des das Enzym kodierende Gen in unterschiedlichen Bakterien. Berechnet wurde dieses Profil auf der Basis der Datenbank KEGG [84]. Hierzu wurden die Informationen über die Existenz der untersuchten Gene in allen Bakterien extrahiert und das Vorkommen der einzelnen Gene in den unterschiedlichen Bakterien gezählt. Ein Gen, das in vielen verschiedenen Bakterien

nachweisbar ist, hat folglich einen hohen Profilwert. Auf der Basis der ebenfalls in KEGG gespeicherten Reaktionsgleichungen wurde dann ein organismenunabhängiges Enzym-Netz erstellt. Für dieses Netz wurden die drei Zentralitäten *Degree*, *Closeness* und *Shortest-Path Betweenness* berechnet und die Zentralitätswerte der einzelnen Enzyme mit dem phylogenetischen Profil desselben verglichen. Es wurde hierdurch gezeigt, dass ein hoher Wert für die *SP-Betweenness-* und die *Degree*-Zentralität mit dem phylogenetischen Profil eines Enzyms positiv korreliert. Für die *Closeness*-Zentralität ist diese Korrelation hingegen schwach negativ. Enzyme bzw. deren kodierenden Gene mit hohem *SP-Betweenness* bzw. *Degree* Wert haben nach dieser Untersuchung folglich eine größere Wahrscheinlichkeit, in verschiedenen Bakterien vorzukommen.

Eine ähnliche Untersuchung haben Lu und Koautoren und Vitkup und Koautoren auf der Basis einer Rekonstruktion des Reaktionsnetzes von *S. cerevisiae* durchgeführt [109, 166]. In den Studien wurden die Zentralitäten *Degree* (Vitkup *et al.*) bzw. *Shortest-Path Betweenness* (Lu *et al.*) verwendet, um eine Aussage über die mögliche evolutionäre Entwicklung von Enzymen bzw. deren kodierender Gene zu machen. Für beide Zentralitäten gilt, dass ein hoher Zentralitätswert auf eine langsamere evolutionäre Entwicklung für das betreffende Gen schließen lässt. Für *E. coli* haben Hahn und Koautoren hingegen keine statistisch signifikante Korrelation zwischen dem *Degree* eines Enzyms und seiner evolutionären Entwicklung gefunden [68].

Wunderlich & Mirny haben ein Maß zur Vorhersage von letalen Mutationen von Enzymen bzw. deren kodierender Gene vorgeschlagen [175]. Dieses *Synthetic Accessibility* genannte Maß basiert auf der Topologie des zu untersuchenden Reaktionsnetzes und liefert einen booleschen Wert, ob nach der Entfernung eines Gens die Lebensfähigkeit des Organismus als gefährdet anzunehmen ist. Als Vergleich werden die drei Zentralitäten *Degree*, *Closeness* und *Shortest-Path Betweenness* herangezogen. Für die drei Zentralitäten ergibt sich laut Wunderlich & Mirny, dass ein hoher Zentralitätswert nicht sicher auf die Lebensnotwendigkeit des Enzyms schließen lässt.

Die im Abschnitt 2.3.5 beschriebene Zentralität *Avalanche* ist nur für bipartite Graphen definiert. Bei dieser Zentralität wird explizit davon ausgegangen, dass der zu analysierende Graph einen Produktionsprozess modelliert. Der Zentralitätswert des betrachteten Knotens (Prozesses) bestimmt sich bei dieser Zentralität anhand der Anzahl der Prozesse bzw. Materialien, die nach der Deaktivierung des betrachteten Prozesses nicht mehr ablaufen bzw. produziert werden können. Durch die Festlegung, dass Reaktionen als Prozesse und Metaboliten als Materialien aufgefasst werden, lässt sich die Zentralität für metabolische Reaktionsnetze einsetzen. In der von Lemke und Koautoren unter dem Namen *Damage* vorgestellten Variante dieser Zentralität werden dabei die Anzahl der nicht mehr produzierbaren Metaboliten gezählt [104]. Ghim und Koautoren hingegen haben die Zentralität auf der Basis der Anzahl der nicht mehr funktionsfähigen Reaktionen unter dem Namen *Avalanche* (deutsch Lawine) eingeführt [62]. Für beide Zentralitäten wurde gezeigt, dass aus einem hohem *Damage-* bzw. *Avalanche*-Wert für ein Enzym eher darauf geschlossen werden kann, dass das Enzym lebensnotwendig für *E. coli* ist. Lemke *et al.* haben zusätzlich noch die Top 30 Liste der Enzyme mit dem höchsten *Damage*-Wert analysiert. Dabei haben sie unter anderem festgestellt, dass zwei Enzyme mit hohem *Damage*-Wert in der Synthese der Zellwand, einem bekannten Angriffspunkt derzeitig verwendeter Antibiotika, eine entscheidende Rolle spielen.

Durch die Anwendung der Zentralität *Load Point* (siehe Def. 2.24) auf den zu einem metabolischen Reaktionsnetz zugehörigen unipartiten Reaktionsgraphen kann eine Reihenfolge der Reaktionen bzw. Enzyme erstellt werden. Als möglichen Einsatzzweck für die-

se Bewertung schlagen Rahman & Schomburg die Erstellung einer Reihenfolge der *Choke Point*-Enzyme vor. *Choke Point*-Enzyme (deutsch Nadelöhr-Enzyme) sind Enzyme, die als einziges eine Reaktion katalysieren, die exklusiv ein Metabolit produziert bzw. verbraucht. In der Kombination mit den Zentralitätswerten der *Load Point*-Zentralität lassen sich die *Choke Point*-Enzyme dann in eine Reihenfolge bringen. Die von Rahman & Schomburg untersuchten Top 10 Enzyme des Bakterium *Bacillus antracis Sterne* zeigen dabei interessante Eigenschaften: zwei Enzyme in *B. antracis Sterne* können als mögliche Ziele für die Medikamentenentwicklung angesehen werden, da sie *Choke Point*-Enzyme für *B. antracis Sterne* allerdings nicht für den Menschen (*H. sapiens*) sind.

Holme und Koautoren haben ein Verfahren zur hierarchischen Zerlegung von (bipartiten) metabolischen Reaktionsnetzen in Teilnetze vorgestellt [72]. Für die Zerlegung des Netzes wird die *Shortest-Path Betweenness*, in diesem Fall nur für die Knotenmenge der Reaktionen berechnet, verwendet. Nach der Berechnung der Zentralitätswerte wird die Reaktion mit dem höchsten Zentralitätswert entfernt und das Netz somit in immer kleinere Komponenten zerlegt. Es werden dabei zwei Arten von Zerlegungen beobachtet: „*Shell-like*“, immer ein kleiner Teil (einige Knoten) werden abgespalten und der Rest bleibt als große Komponente erhalten und „*Community-type ordering*“, das Netzwerk zerfällt in mehrere ähnlich große Komponenten.

4.2.3. Vergleich von Netzen auf der Basis von Zentralitätsmaßen

Viele Autoren nutzen Zentralitätsmaße zum Vergleich der Struktur von Netzwerken. Das einfachste Beispiel ist hierbei die Gradverteilung. Sie basiert auf dem *Degree* und gibt an, wie viele Knoten mit einem bestimmten Grad im Graphen vorkommen. Eingesetzt wurde die Gradverteilung zur Beschreibung metabolischer Reaktionsnetze beispielsweise von Jeong [79] und Wagner & Fell [167]. Zhu & Qin wiederum nutzten die *Shortest-Path Betweenness* (und weitere Maße zur Charakterisierung von Netzen) um die Entwurfsprinzipien der Reaktionsnetze der drei Domänen (Bakterien, Archaeen und Eukaryoten) zu vergleichen [180]. Die Zentralitätswerte werden dabei nur als Mittelwert genutzt. Zhu und Qin schlussfolgern aus ihrer Analyse, dass die Netze der Organismen aus der Domäne der Archaeen sich deutlich von den Netzen der Organismen der anderen beiden Domänen unterscheiden.

4.3. Publizierte Zentralitätsanalysen von Protein-Interaktionsnetzen

Eine Reihe von Protein-Interaktionsnetzen wurden bereits mit Hilfe von Zentralitäten untersucht. Der Fokus der Analysen lag dabei in der Begründung bzw. Widerlegung eines Zusammenhangs zwischen der Bewertung eines Knotens und verschiedener funktioneller Eigenschaften der im Netz modellierten Proteine bzw. deren Interaktionen. Hierbei kamen die Zentralitätsmaße *Degree*, *Eccentricity*, *Closeness*, *CF-Closeness*, *Subgraph*, *Bipartivity*, *Transitivity*, *SP-Betweenness*, *CF-Betweenness*, *Eigenvector*, *Entropy* und *Damage* zum Einsatz. Im Folgenden (siehe auch Tabelle 4.3) werden die Resultate der bisher publizierten Analysen zusammengefasst.

4.3.1. Vorhersage der Lebensnotwendigkeit von Proteinen

Ob ein Protein für einen Organismus lebensnotwendig ist oder ob der Organismus auch ohne das betreffende Protein überleben kann ist insbesondere im Bereich der Medikamentenent-

wicklung von großem Interesse. Ein lebensnotwendiges Protein für ein Bakterium, welches gleichzeitig beispielsweise für den Menschen nicht lebensnotwendig ist, könnte ein möglicher Angriffspunkt gegen das Bakterium sein. Dieses Protein käme somit als potentielles Ziel während der Entwicklung eines Antibiotikums in Frage. Die Unterscheidung zwischen lebensnotwendigen und nicht lebensnotwendigen Proteinen auf der Basis der Struktur des Protein-Interaktionsnetzes wurde deshalb bereits vielfältig untersucht. Hierzu wurden, neben anderen Verfahren, auch Zentralitäten eingesetzt. Es wird dabei eine Reihenfolge der Proteine auf der Basis der Zentralitätswerte erstellt und angenommen, dass Proteine mit einem hohen Zentralitätswert eher lebensnotwendig sind und im Gegensatz der Organismus bei Verlust eines Proteins mit einem geringeren Zentralitätswert wahrscheinlich trotzdem lebensfähig ist.

Die erste Veröffentlichung, in der dieser Zusammenhang untersucht wurde, wurde 2001 von Jeong und Koautoren publiziert. In dieser Analyse wurde gezeigt, dass eine positive Korrelation zwischen einem hohen *Degree* und der Wahrscheinlich für die Lebensnotwendigkeit des betreffenden Proteins im Protein-Interaktionsnetz von *S. cerevisiae* existiert [77].

In einer ganzen Reihe von nachfolgenden Publikationen wurde derselbe Zusammenhang im mehr oder weniger starker Form, jeweils für andere Protein-Interaktionsnetze[6] der *S. cerevisiae*, bestätigt [70, 134, 137, 143, 173, 177].

Zur Verbesserung der Vorhersagekraft schlugen Jeong und Koautoren in einer nachfolgenden Veröffentlichung vor weitere Informationen in die Berechnung der Wahrscheinlichkeit für die Lebensnotwendigkeit mit aufzunehmen. Es wurde empfohlen, Informationen über die Zuordnung der Proteine zu funktionellen Gruppen und Messergebnissen für den jeweiligen mRNA-Expressionslevel des Proteins mit in die Berechnung einfließen zu lassen [78].

Joy *et al.* und Yu *et al.* haben für zwei unterschiedliche Protein-Interaktionsnetze von *S. cerevisiae* gezeigt, dass die für die *Degree*-Zentralität beobachtete Korrelation auch für die *Shortest-Path Betweenness* gilt [80, 178].

Die generelle Eignung von Zentralitäten zur Vorhersage von lebensnotwendigen Proteinen haben Hahn & Kern und Estrada untersucht. Für Protein-Interaktionsnetze von *S. cerevisiae*, *D. melanogaster* und *C. elegans* und die Zentralitäten *Degree*, *Closeness* und *Shortest-Path Betweenness* ist jeweils der durchschnittliche Zentralitätswert für lebensnotwendige Proteine höher als der entsprechende Wert für nicht lebensnotwendige Proteine [69]. Für ein Protein-Interaktionsnetz von *S. cerevisiae* wurden die Zentralitäten *Degree*, *Closeness*, *Shortest-Path Betweenness*, *Eigenvector*-Zentralität, *Information Centrality*, *Subgraph*-Zentralität und *Bipartivity* mit der zufälligen Auswahl von Proteinen für die Bestimmung der lebensnotwendigen Proteinen verglichen. Bei diesem spezifischen Netz ergibt sich eine Reihenfolge der Eignung der Zentralitäten. Diese Reihenfolge wird von den Zentralitäten *Bipartivity* und *Subgraph*-Zentralität angeführt [45, 46].

Die Zentralitäten *Damage*, *Entropy* und *Transitivity* wurden ebenfalls zur Bestimmung lebensnotwendiger Proteine eingesetzt. Die Zentralität *Damage* ist für die Erkennung lebensnotwendiger Proteine geeignet. Sie schneidet allerdings im direkten Vergleich mit der *Degree*-Zentralität etwas schlechter ab, da in vier von fünf untersuchten Protein-Interaktionsnetzen für *S. cerevisiae* die Korrelation zwischen dem *Degree*-Wert und der Eigenschaft „lebensnotwendig“ höher ist als zwischen dem *Damage*-Wert und derselben Eigenschaft [151]. Für die Zentralität *Entropy* gilt, dass diese für je ein Netz für die Organismen *S. cerevisiae* und *C. elegans* zur Unterscheidung von lebensnotwendigen von nicht lebensnotwendigen Proteinen geeignet ist [118, 119]. Und für ein Protein-Interaktionsnetz von *S. cerevisiae* gilt ebenso,

[6] Unterschiedliche Netzwerke entstehen bei Verwendung unterschiedlicher Datenquellen.

dass ein höherer Wert für die Zentralität *Transitivity* eher auf ein lebensnotwendiges Protein schließen lässt [172].

In einige Analysen wurden keine bzw. negative Zusammenhänge zwischen hohen Zentralitätswerten und der Wahrscheinlichkeit für das Vorliegen eines lebensnotwendigen Proteins gefunden [36, 71, 174]. Coulomb und Koautoren nennen als mögliche Begründung, dass bei der Betrachtung der Protein-Interaktionsnetze das bei der Erstellung der Netze mit einfließende Bias nicht ausreichend berücksichtigt wird. Beispielsweise wird bei der Erstellung von Protein-Interaktionsnetzen auf der Basis der vorhandenen Literatur die Tatsache, dass wahrscheinlich mehr Publikationen zu lebensnotwendigen Proteinen gegenüber nicht lebensnotwendigen Proteinen existieren, nicht hinreichend berücksichtigt.

4.3.2. Evolutionsbedingte Veränderungen in der Proteinsequenz

In mehreren Analysen wurde der Zusammenhang zwischen der Position eines Proteins innerhalb eines Protein-Interaktionsnetzes mit der Anzahl der Unterschiede in der Proteinsequenz eines vergleichbaren Proteins in einem biologisch nahestehenden Organismus verglichen. Hierzu wurde jeweils zum untersuchten Protein ein entsprechendes (orthologes) Protein aus dem Vergleichsorganismus bestimmt und die Proteinsequenz der beiden miteinander verglichen.

Für unterschiedliche Protein-Interaktionsnetze von *S. cerevisiae* und unter Berücksichtigung unterschiedlicher Vergleichsorganismen wurde dabei ein negativer Zusammenhang zwischen den beiden Zentralitäten *Degree* und *Shortest-Path Betweenness* auf der einen Seite und der Anzahl der Änderungen an der Proteinsequenz auf der anderen Seite gezeigt. Dieser negative Zusammenhang verdeutlicht, dass Proteine mit einem hohem Zentralitätswert im Verhältnis weniger evolutionsbedingte Änderungen an ihrer Proteinsequenz erfahren als Proteine mit einem niedrigerem Zentralitätswert [56, 68, 80, 173].

In einem größeren Vergleich wurde derselbe Zusammenhang für die Zentralitäten *Degree*, *Closeness* und *Shortest-Path Betweenness* in jeweils einem Protein-Interaktionsnetz von *S. cerevisiae*, *D. melanogaster* und *C. elegans* untersucht. Für die *Degree* und die *Shortest-Path Betweenness* ist derselbe Zusammenhang wie oben sichtbar, für die *Closeness*-Zentralität hingegen nicht [69].

4.3.3. Weitere Anwendungen von Zentralitäten auf Proteininteraktionsnetze

Im einem Protein-Interaktionsnetz von *S. cerevisiae* wurde beobachtet, dass Proteine mit hohem *Degree* (*Hubs*) mit hoher Wahrscheinlichkeit auf genetischer Ebene[7] mit einem anderen *Hub*-Protein interagieren [130]. Eine mögliche genetische Interaktionen ist dabei die gemeinsame Lebensnotwendigkeit beider Proteine. Wenn eines der beiden Proteine deaktiviert wird, dann ist die betreffende Zelle weiterhin lebensfähig. Wenn allerdings beide Proteine gleichzeitig deaktiviert werden, dann ist die betreffende Zelle nicht lebensfähig.

Die Eigenschaften von *Hub*-Proteinen aus unterschiedlichen Protein-Interaktionsnetzen von *S. cerevisiae* wurde von Batada und Koautoren untersucht. Unter anderem wurde der Zusammenhang zwischen der Lebensnotwendigkeit und dem hohen *Degree*-Wert eines Proteins dabei erneut bestätigt [18].

[7] Eine genetische Interaktion ist eine Beziehung zwischen zwei Genen, bei denen die Mutation jeweils eines Gens keine (beobachtbare) Veränderung hervorruft und die gemeinsame Veränderung zu einem beobachtbaren Effekt führt.

Hub-Proteine haben (definitionsgemäß) viele Interaktionen. Die unterschiedlichen Interaktionen zwischen je zwei Proteinen müssen allerdings innerhalb der Zelle zeitlich oder räumlich nicht zusammenhängen. Die Unterscheidung zwischen Interaktionen, die zeitgleich und Interaktionen, die eher nacheinander stattfinden können, führte zur Definition von *Party Hubs* und *Date Hubs*. Für *Party Hubs* finden die Interaktionen nahezu zeitgleich statt. Interaktionen bei den *Date Hubs* hingegen finden eher zu unterschiedlichen Zeitpunkten statt [70].

Für ein Protein-Interaktionsnetz von *H. sapiens*, in dem nur die Transkriptionsfaktoren und ihre Interaktionen modelliert wurden, wurden die Zentralitäten *Degree* und *Shortest-Path Betweenness* zur Erstellung einer Reihenfolge der Transkriptionsfaktoren genutzt. Auf der Basis dieser Ordnung wurden die Top 9 Transkriptionsfaktoren bzgl. der *Degree*-Zentralität identifiziert und beschrieben. Dominiert wird diese Liste von Transkriptionsfaktoren, die mit der Krankheit Krebs in Verbindung stehen [144].

Koschützki und Schreiber haben fünf Zentralitäten (*Degree*, *Eccentricity*, *Closeness*, *Random-Walk Betweenness* und *Eigenvector*-Zentralität) auf ein Protein-Interaktionsnetz von *H. sapiens* angewendet [97]. Dabei wurde, wie auch für ein Genregulationsnetz für *E. coli*, beobachtet, dass die einzelnen Zentralitätsmaße miteinander korrelieren. Im Unterschied zum Genregulationsnetz korrelieren beim Protein-Interaktionsnetz die Zentralitäten *Eigenvector* und *Eccentricity* stark miteinander ($r = 0,9248$, Pearsonscher Maßkorrelationskoeffizient [93]).

4.4. Schlussfolgerungen aus den bisherigen Publikationen

Auf der Basis der oben zusammengefassten bisher publizierten Zentralitätsanalysen von molekularbiologischen Netzwerken wird jetzt die Notwendigkeit neuer, an den entsprechenden Netzwerktyp angepasster Zentralitätsmaße motiviert.

4.4.1. Genregulationsnetze

Bisher sind erst zwei Analysen von Genregulationsnetzen mit Hilfe von Zentralitäten publiziert worden. In beiden Fällen waren die angewendeten Zentralitätsmaße bereits vorher (aus der Analyse sozialer Netzwerke) bekannt. Eine spezielle Motivation, warum die angewendeten Zentralitätsmaße zur Analyse der Genregulationsnetze geeignet sind, erfolgte in beiden Publikationen nicht. Im Kapitel 5 wird deshalb erstmalig eine Zentralität zur Analyse von Genregulationsnetzen vorgestellt.

4.4.2. Metabolische Reaktionsnetze

Die bisher zur Analyse metabolischer Reaktionsnetze eingesetzten Zentralitäten basieren auf insgesamt vier Prinzipien: Nachbarschaft (*Degree*), Wege (*Eccentricity*, *Closeness*, *Centroid Value*, *Load Points*), Beobachtbarkeit (*Shortest-Path Betweenness*) und Anzahl deaktivierter Netzelemente (*Damage*, *Avalanche*). Die beiden Zentralitäten des vierten Prinzips (Anzahl deakt. Netzwerkelemente) nutzen eine spezifische Eigenschaft metabolischer Reaktionsnetze zur Bewertung aus: wenn eines der Metaboliten dem Organismus nicht zur Verfügung steht, dann können alle davon abhängigen Reaktionen nicht ablaufen und folglich stehen weitere Metaboliten bzw. Reaktionen nicht zur Verfügung.

Die durch die Verwendung der entsprechenden Zentralitäten (*Eccentricity*, *Closeness*, *Centroid Value*, *Load Points*, *Shortest-Path Betweenness*) getroffene Annahme, dass meta-

bolische Reaktionen ausschließlich über kürzeste Wege ablaufen, ist anzuzweifeln. Innerhalb eines Organismus findet die „Auswahl“ der ablaufenden Reaktionen auf der Basis vieler Einflüsse, beispielsweise der Stoffkonzentration, der Temperatur und dem in der Zelle vorherrschenden pH-Wert, statt. Hierbei kann durchaus eine Kette von Reaktionen aktiv sein, die nicht dem kürzesten Weg zwischen zwei Metaboliten entspricht. Die Analyse metabolischer Reaktionsnetze mittels Zentralitäten, die kürzeste Wege als Bewertungsmaßstab einsetzen, ist folglich nur bedingt möglich.

Die Anzahl der deaktivierten Reaktionen bzw. Metaboliten hingegen ist für die Analyse metabolischer Reaktionsnetze eine sinnvolle Betrachtungsweise. Insbesondere in der Medikamentenentwicklung könnte beispielsweise die Identifikation einer Reaktion mit hohem Zentralitätswert bzgl. der Zentralitäten *Damage* und *Avalanche* von Interesse sein, da es sich hierbei um ein potentielles Ziel beispielsweise für die Unschädlichmachung von Bakterien durch ein Antibiotikum handeln könnte.

Die in einem Organismus durch den Metabolismus umgesetzte Stoffmenge pro Reaktion wird *metabolischer Fluss* genannt. Dieser Fluss hängt von vielen Faktoren, insbesondere den Umgebungsbedingungen des Organismus, ab. Aktuell wird dieser Fluss von keiner Zentralität bei der Bestimmung einer Reihenfolge der Netzwerkelemente verwendet. Durch die Hinzunahme der Informationen über den Fluss lässt sich eine, von den Umgebungsbedingungen abhängige, und der Bedeutung der Metaboliten bzw. Reaktionen besser entsprechende Reihenfolge derselben ermitteln. Im Kapitel 6 wird deshalb eine Zentralität vorgestellt, die die spezifischen Eigenschaften metabolischer Reaktionsnetze, insbesondere den metabolischen Fluss unter unterschiedlichen Wachstumsbedingungen, für die Erstellung einer Reihenfolge der Metaboliten verwendet.

4.4.3. Protein-Interaktionsnetze

Die bisher zur Analyse von Protein-Interaktionsnetzen eingesetzten Zentralitäten basieren auf den folgenden Prinzipien: Nachbarschaft (*Degree*), Wege (*Closeness*, *Subgraph*-Zentralität, *Bipartivity* und *Transitivity*), Beobachtbarkeit (*Shortest-Path Betweenness*), Rückkopplung (*Eigenvector*-Zentralität, *Information Centrality*, *Entropy*) und Anzahl deaktivierter Netzelemente (*Damage*). Der Einsatz der Zentralitäten erfolgte dabei jeweils ohne weitere Begründung für die Auswahl. Eine Betrachtung, welches Prinzip für die Analyse von Protein-Interaktionsnetzen geeignet ist und die entsprechenden biologischen Vorgängen adäquat abbildet, erfolgte bisher weder für die bereits bekannten, noch für die neu vorgestellten Zentralitäten.

Zusätzlich zur Problematik, dass das zu nutzende Zentralitätenprinzip für Protein-Wechselwirkungen noch nicht bekannt ist, besteht bei Protein-Interaktionsnetzen das Problem, dass die zu analysierenden Netze größtenteils experimentell ermittelt sind. Die hierbei eingesetzten Methoden sind teilweise sehr fehleranfällig [176]. Beispielsweise treten bei der relativ einfach einzusetzenden Methode *Yeast-Two Hybrid* eine hohe Anzahl von falsch-positiven Beobachtungen, d.h. Beobachtungen von Interaktionen im Experiment, obwohl in der lebenden Zelle keine entsprechende Interaktion stattfindet, auf [10]. Diese experimentell-bedingten Fehler in der Beobachtung einzelner Interaktionen sind dann auch Bestandteil des aus den einzelnen Interaktionen zusammengesetzten Gesamtnetzes.

Zur Beurteilung der Qualität der erstellten Netze wurden durch unterschiedliche experimentelle Verfahren ermittelte Protein-Interaktionsnetze desselben Organismus miteinander verglichen. Hierbei zeigten sich teilweise deutliche Unterschiede in diesen Netzen. In einer Untersuchung von acht verschiedenen Interaktionsnetzen für *H. sapiens* wurde kein Inter-

aktionspaar in mehr als fünf untersuchten Netzen gleichzeitig beobachtet [58]. Auch für *S. cerevisiae* wurde Vergleichbares berichtet [123]. Die Bewertung der Güte eines Interaktionsnetzes wird durch das Fehlen eines Referenzdatensatzes erschwert. In diesem Datensatz sollten, zusätzlich zu bekannten Interaktionen, auch Informationen über die Nicht-Existenz von Interaktionen enthalten sein [76]. Zur Zeit ist ein solcher als *Gold-Standard* bezeichneter Datensatz allerdings noch für keinen Organismus verfügbar.

Da die Qualität der untersuchten Daten zur Zeit nur bedingt beeinflusst werden kann, ist bei jeder strukturellen Analyse von Protein-Interaktionsnetzen folglich eine Berücksichtigung von möglichen Fehlern in den Daten erforderlich. Diese Fehler in den Daten können beispielsweise durch die Analyse der Stabilität der einzelnen Zentralitätsmaße gegenüber kleineren Änderungen in den Netzen berücksichtigt werden. Die Idee hierbei ist, dass, wenn zur Analyse des Netzes eine Zentralität verwendet wird, die gegenüber kleineren Änderungen innerhalb der Netzwerkstruktur stabil ist, dass dann das berechnete Ranking entsprechend auch für das (nicht bekannte) „wirkliche“ Protein-Interaktionsnetz genutzt werden kann. Zu einigen bereits zur Analyse von Protein-Interaktionsnetzen eingesetzten Zentralitäten wurden bereits Analysen zur Stabilität durchgeführt, beispielsweise für *Degree*, *Closeness*, *Shortest-Path Betweenness* und die *Eigenvector*-Zentralität [35] und für die Zentralität *Entropy* [118, 119]. Bisher wurde allerdings noch keine vertiefende Betrachtung, insbesondere in Hinblick auf die Spezifika von Protein-Interaktionsnetzen, durchgeführt.

Beide Punkte, die Auswahl eines geeigneten Zentralitätenprinzips und die Auswahl einer gegen Änderungen in der Netzwerkstruktur robusten Zentralität, werden im Folgenden nicht weiter vertieft. Hierzu sind noch umfassende Arbeiten im Bereich der Zentralitätenprinzipien für Protein-Interaktionsnetzes, im Messen der Robustheit von Zentralitäten und darauf aufbauend in der Entwicklung eines robusten, für die Analyse von Protein-Interaktionsnetzen geeigneten Zentralitätsmaßes erforderlich.

4.5. Zusammenfassung

In diesem Kapitel wurden die existierenden Zentralitätsanalysen für die betrachteten molekularbiologischen Netzwerke vorgestellt und zusammengefasst. Dabei wurde festgestellt, dass für Genregulationsnetze bisher noch mit Abstand die geringste Anzahl an Analysen durchgeführt wurden, dass für die Analyse metabolischer Reaktionsnetze die Betrachtung des metabolischen Flusses bei der Berechnung der Zentralitätswerte berücksichtigt werden sollte und dass Zentralitäten für Protein-Interaktionsnetze hinreichend robust bzgl. Änderungen in der Netzstruktur, bedingt durch unvollständige bzw. fehlerhaft ermittelte Daten, sein sollten.

In den beiden folgenden Kapiteln wird jeweils ein Zentralitätsmaß für die Netzwerktypen Genregulationsnetz und metabolisches Reaktionsnetz vorgestellt. Die Zentralität zur Analyse von Genregulationsnetzen basiert dabei auf der Idee der Netzwerkmotive und die Zentralität für die Bestimmung einer Reihenfolge der Metaboliten nutzt Informationen über den metabolischen Fluss.

Zentralität	Organismenübergr.	*E. coli*	*S. cerevisiae*	(a)	(b)	(c)	(d)	(e)
Degree	[106, 79]	[14, 50, 49, 68, 114, 167, 175]	[79, 114, 166]	—	—	[114]	—	
Eccentricity	—	[174]	—	—	—	—	—	—
Closeness	[106]	[50, 49, 110, 167, 174, 175]	—	—	—	—	—	
Centroid-Value	—	[174]	—	—	—	—	—	—
Load Point	—	—	—	[140]	[140]	—	—	—
CF-Closeness	—	—	—	—	—	—	—	—
Subgraph	—	—	—	—	—	—	—	—
Bipartivity	—	—	—	—	—	—	—	—
Transitivity	—	—	—	—	—	—	—	—
SP-Betweenness	[106, 180]	[175]	[109]	—	—	—	[72]	[72]
CF-Betweenness	—	—	—	—	—	—	—	—
Eigenvector	—	—	—	—	—	—	—	—
Entropy	—	—	—	—	—	—	—	—
Damage	—	—	—	—	—	—	—	—
Avalanche	—	[62, 104]	—	—	—	—	—	—

(a) *Bacillus antracis St.*
(b) *Bacillus subtilis 168*
(c) *Geobacter sulfurreducens*
(d) *Treponema pallidum*
(e) *Mycoplasma pneumoniae*

Tabelle 4.2.: Aufstellung existierender Publikationen, in denen metabolische Reaktionsnetze mittels Zentralitäten analysiert wurden. In den Zeilen sind die in diesem Dokument beschriebenen Zentralitäten aufgeführt und in den Spalten die Namen der Organismen, die in den benannten Publikationen analysiert wurden.

Zentralität	*S. cerevisiae*	*H. sapiens*	*C. elegans*	*D. melanogaster*
Degree	[18, 36, 45, 46, 56, 68, 69, 70, 71, 77, 78, 80, 130, 134, 137, 143, 173, 174, 177]	[97, 144]	[69]	[69]
Eccentricity	—	[97]	—	—
Closeness	[45, 46, 69, 71]	[97]	[69]	[69]
Centroid-Value	—	—	—	—
Load Point	—	—	—	—
CF-Closeness	[45, 46]	—	—	—
Subgraph	[45, 46]	—	—	—
Bipartivity	[45]	—	—	—
Transitivity	[172]	—	—	—
SP-Betweenness	[45, 46, 69, 71, 80, 178]	[144]	[69]	[69]
CF-Betweenness	—	[97]	—	—
Eigenvector	[45, 46]	[97]	—	—
Entropy	[118, 119]	—	[118, 119]	—
Damage	[151]	—	—	—
Avalanche	—	—	—	—

Tabelle 4.3.: Aufstellung existierender Publikationen, in denen Proteininteraktionsnetze mittels Zentralitäten analysiert wurden. In den Zeilen sind die in diesem Dokument beschriebenen Zentralitäten aufgeführt und in den Spalten die Namen der Organismen, die in den benannten Publikationen analysiert wurden.

5. Motiv-basierte Zentralitäten für die Analyse von Genregulationsnetzen

In Netzwerken unterschiedlichen Typs, beispielsweise Netzwerken, die Routerverbindungen beschreiben, Modellen der Verlinkung von Webseiten („Webgraphen“), digitalen Schaltkreisen oder auch molekularbiologischen Netzwerken, wurden kleine Teilnetze (engl. *Subgraphs*, *Patterns* oder *Motifs*) identifiziert, von denen vermutet wird oder bereits bekannt ist, dass diese eine funktionelle Bedeutung für das modellierte System haben [125, 158].

Die bereits bekannten Zentralitätsmaße berücksichtigen allerdings die Informationen über das Auftreten und die Häufigkeit des Auftretens dieser Teilnetze nicht. Ein Zentralitätswert wird bei diesen Zentralitätsmaßen entweder auf der Basis der unmittelbaren Nachbarschaft des Knotens (*Degree* Zentralität, siehe Seite 10) oder über die Betrachtung der gesamten Struktur des Netzwerkes, wie beispielsweise bei der *Closeness* Zentralität (siehe Seite 12) oder der *Eigenvector* Zentralität (siehe Seite 17), berechnet.

Im Folgenden wird deshalb eine Klasse von Zentralitätsmaßen vorgestellt, die es erlaubt Teilnetze zur Bewertung von Knoten eines zu untersuchenden Netzes zu verwenden. Es wird dabei davon ausgegangen, dass ein Knoten, der in vielen verschiedenen Instanzen eines gesuchten Teilnetzes auftritt eine besonders große Rolle innerhalb des untersuchten Netzwerkes spielt. Folglich soll dieser Knoten einen entsprechend höheren Zentralitätswert im Vergleich zu anderen Knoten erhalten.

Da Motive im Genregulationsnetzwerk von *E. coli* bereits sehr gut untersucht sind, werden die hier vorgestellten Zentralitätsmaße auf ein Genregulationsnetzwerk von *E. coli* angewendet. Dabei werden interessante Beobachtungen über die Struktur des Netzes gemacht und eine Liste von globalen Regulatoren ermittelt. Abschließend werden die Resultate der Analyse auf der Basis der neuen Zentralitäten mit Analyseresultaten auf der Basis der bereits existierenden Zentralitäten, die ebenfalls auf Genregulationsnetze angewendet werden können, verglichen.

5.1. Motive in Netzwerken

Aus der Sequenzanalyse ist der Begriff des (Sequenz-) Motivs bekannt. Hierbei handelt es sich um einen Abschnitt einer DNA- oder Protein-Sequenz der in mehreren Ausprägungen, eventuell in leicht veränderter Form, auftritt. Von einigen dieser Sequenzmotiven ist die biologische Funktion bekannt. Aus dem Vorkommen eines bekannten Motivs in der Sequenz eines noch nicht weiter untersuchten Proteins kann somit evtl. auf die mögliche Funktion des Proteins geschlossen werden [107, 113].

In Anlehnung an den Begriff des Sequenzmotivs werden Teilnetze die gegenüber einer Randomisierung desselben Netzes signifikant häufiger auftreten ebenfalls *Motive* (engl. *Motif*) genannt [125, 158]. Wie für die Sequenzmotive wird auch bei den (Netzwerk-) Motiven[1]

[1] Im Folgenden bezeichnet der Begriff „Motiv“ immer ein Netzwerkmotiv.

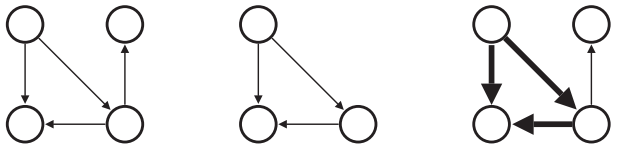


Abbildung 5.1.: (a) Ein Quellgraph G, (b) ein Motiv M und (c) ein Treffer G_M von M in G.

in molekularbiologischen Netzwerken davon ausgegangen, dass diese eine funktionelle Bedeutung haben [9, 116, 117].

Bisher wurden Netzwerkmotive beispielsweise in Schaltkreisen, einem Teil des Webgraphen und einem Netzwerk, das Verbindungen zwischen Routern modelliert identifiziert [125]. In biologischen Netzwerken wurden Motive in Genregulationsnetzen [85, 158], einem Protein-Interaktionsnetz [171] und einer Nahrungskette [125] gefunden.

Für die Bestimmung der signifikant häufiger auftretenden Teilgraphen in einem zu untersuchenden Netzwerk existieren eine Reihe von Algorithmen [152, 158, 160] und Tools [86, 153, 170]. Das Resultat der Analyse eines zu untersuchenden Netzwerkes mit Hilfe dieser Verfahren ist üblicherweise eine Liste von Teilgraphen einer bestimmten Größe, beispielsweise der Teilgraphen mit vier Knoten, die in dem Netz auftreten und die zugehörigen Häufigkeiten.

Für die im Folgenden zu beschreibenden Zentralitäten kann jeder Teilgraph eines zu untersuchenden Netzwerks verwendet werden. Die Auswahl des am häufigsten auftretenden Motivs ist nicht erforderlich. Allerdings ändern sich durch die Verwendung unterschiedlicher Teilgraphen die Zentralitätswerte und die zugrunde liegende Bedeutung bzw. Interpretation der Zentralitätswerte.

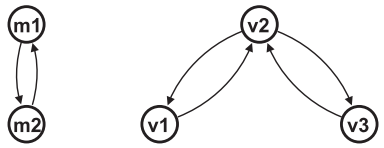
Formal ist ein Motiv ein Graph. Wenn dieser Graph in einem zu untersuchenden Graphen („Quellgraph“) als Teilgraph (siehe Definition 2.4) auftritt, dann wird dieser Teilgraph des Quellgraphen *Treffer* (engl. *Match*) des Motivs im Quellgraphen genannt.

Definition 5.1 (Treffer)
Sei $G = (V_G, E_G)$ ein gerichteter Graph[2], der Quellgraph*, und sei $M = (V_M, E_M)$ ein gerichteter Graph, das* Motiv*. Ein* Treffer *G_M des Motivs M im Quellgraph G ist ein Teilgraph von G ($G_M \subseteq G$) der isomorph zum Motiv M ($G_M \simeq M$) ist.*

Die Abbildung 5.1 zeigt einen Quellgraphen, ein Motiv und den einzigen Treffer des Motivs innerhalb des Quellgraphen. In der obigen Definition des Motivtreffers wird gefordert, dass der Treffer G_M ein Teilgraph des untersuchten Graphen G ist. Nicht gefordert ist an dieser Stelle die Eigenschaft „induzierter Teilgraph“ (Definition 2.4). Diese Unterscheidung ist wesentlich, da die strengere Definition „induzierter Teilgraph“ die Menge der gefundenen Treffer zu stark einschränkt.

Die Gesamtmenge aller Treffer eines Motivs in einem Quellgraphen wird *Treffermenge* genannt.

[2] Alle Definitionen und Aussagen in diesem Kapitel basieren auf gerichteten Graphen. Im Rahmen der Motiv-basierten Zentralitäten können ungerichtete Graphen durch gerichtete Graphen mit antiparallelen Kanten repräsentiert werden.



(a) (b)

	M_1	M_2	M_3	M_4
m_1	v_1	v_2	v_2	v_3
m_2	v_2	v_1	v_3	v_2

(c)

Abbildung 5.2.: (a) Ein Motiv, (b) ein Quellgraph, (c) die Knotenmengen der vier Treffer zwischen dem Motiv und dem Quellgraph.

Definition 5.2 (Treffermenge)
Sei G ein Graph und sei M ein Motiv (wie in Def. 5.1). Die Menge aller Treffer des Motivs M im Graphen G wird Treffermenge $\mathcal{G}_M$ *genannt. Sie ist definiert als* $\mathcal{G}_M := \{G_M \mid G_M \subseteq G \wedge G_M \simeq M\}$.

Die Treffermenge $\mathcal{G}_M$ ist somit eine Menge von Teilgraphen von G.

Für die im Folgenden zu beschreibenden Motiv-basierten Zentralitäten bildet die Treffermenge die Grundlage. Hierbei ist zu beachten, dass Teilgraphen, die als Treffer auftreten nur einmal in der Treffermenge enthalten sind, auch wenn es mehrere Teilgraph-Isomorphismen gibt. Abbildung 5.2 verdeutlicht diesen Effekt: Das Motiv (a) tritt insgesamt viermal im Quellgraphen (b) auf, d.h. es existieren vier Isomorphismen (c) zwischen der Knotenmenge des Motivs und einer Teilmenge der Knotenmenge des Quellgraphen. Zu diesen vier Isomorphismen gehören allerdings nur zwei unterschiedliche Teilgraphen von G: $G_1 = (\{v_1, v_2\}, E(G) \cap (\{v_1, v_2\} \times \{v_1, v_2\}))$ und $G_2 = (\{v_2, v_3\}, E(G) \cap (\{v_2, v_3\} \times \{v_2, v_3\}))$. Nur diese beiden Graphen sind in der Treffermenge des Motivs M bzgl. des Quellgraphen G enthalten.

5.2. Zentralitäten auf der Basis von Netzwerkmotiven

Auf der Basis einer gegebenen Treffermenge $\mathcal{G}_M$ eines Motivs M in einem Graphen G lassen sich drei Zentralitätsmaße bzw. Familien von Zentralitätsmaßen definieren: Die Motiv-basierte Zentralität, die Rollen-basierte Motiv-basierte Zentralitätsfamilie und die Rollen-basierte Motiv-basierte Zentralitätsfamilie für Klassen von Motiven [98, 99]. Im Folgenden werden diese drei Zentralitätsmaße vorgestellt.

5.2.1. Motiv-basierte Zentralität

Die Motiv-basierte Zentralität ist die einfachste Variante der hier vorgestellten Zentralitäten. Bei dieser Zentralität fließt jeder gefundene Treffer eines Motivs gleich gewichtet in die Berechnung der Zentralitätswerte mit ein.

Definition 5.3 (Motiv-basierte Zentralität)
Sei $\mathcal{G}_M$ die Treffermenge eines Motivs $M = (V_M, E_M)$ zu einem Graphen $G = (V_G, E_G)$. Die Funktion $\mathcal{C}_{mc}^M \colon V_G \mapsto \mathbb{R}$ definiert als $\mathcal{C}_{mc}^M(v) := |\{G_M \mid G_M \in \mathcal{G}_M \wedge v \in V(G_M)\}|$ ist eine Zentralität und wird Motiv-basierte Zentralität (basierend auf dem Motiv M) *genannt.*

Bei der Motiv-basierten Zentralität bestimmt sich der Zentralitätswert eines Knotens $v \in V_G$ somit aus der Anzahl der Treffer (aus der Treffermenge $\mathcal{G}_M$), die den Knoten v enthalten.

(a) (b)

Abbildung 5.3.: (a) Das Motiv Feed-Forward Loop (FFL) und (b) ein Beispielgraph.

Knoten	Zentralitätswert $\mathcal{C}_{mc}^{\text{FFL}}$
v_1	1
v_2	3
v_3	2
v_4	2
v_5	1

Tabelle 5.1.: Die Zentralitätswerte für den Graphen aus Abbildung 5.3(b) berechnet mit der Motiv-basierten Zentralität für das Feed-Forward Loop Motiv (Abbildung 5.3(a)).

Die Abbildung 5.3 zeigt auf der linken Seite ein Motiv, das Feed-Forward Loop Motiv (FFL) [116, 117], und auf der rechten Seite einen Quellgraphen. Innerhalb dieses Graphen tritt das FFL Motiv dreimal auf. Die Treffermenge besteht aus drei Teilgraphen des Quellgraphen und die Knotenmengen dieser Treffer lauten: $\{v_1, v_2, v_3\}$, $\{v_2, v_3, v_4\}$ und $\{v_2, v_4, v_5\}$. In der Tabelle 5.1 sind die Zentralitätswerte für die fünf Knoten des Quellgraphen dargestellt. Da der Knoten v_2 als einziger Knoten in allen drei Treffern auftritt, erhält dieser den höchsten Zentralitätswert.

5.2.2. Rollen-basierte Motiv-basierte Zentralitätsfamilie

Innerhalb eines Motivs können die einzelnen Knoten unterschiedliche Funktionen repräsentieren. Im FFL Motiv aus Abbildung 5.3(a) existieren beispielsweise drei unterschiedliche Funktionen: Der oben stehende Knoten übt Einfluss auf die beiden anderen Knoten aus, der Knoten rechts wird durch den Knoten oben kontrolliert und übt selbst Einfluss auf den Knoten unten aus und der untere Knoten hat selbst keinen Einfluss, wird aber durch die beiden anderen Knoten kontrolliert. Die unterschiedlichen Funktionen der Knoten innerhalb eines Motivs werden im Folgenden *Rollen* genannt.

Bei der Zuweisung von Rollen an die Knoten des betrachteten Motivs sind insgesamt drei Bedingungen zu beachten: a. jedem Knoten des Motivs muss genau eine Rolle zugewiesen werden, b. die den Knoten zugewiesenen Rollen können, müssen aber nicht, paarweise verschieden sein, d.h. dieselbe Rolle kann mehreren Knoten zugewiesen werden, und c. allen Knoten mit identischem Orbit (siehe Seite 6) muss dieselbe Rolle zugewiesen werden.

Basierend auf der Idee der Rollen kann eine Familie von Zentralitäten definiert werden:

Definition 5.4 (Rollen-basierte Motiv-basierte Zentralitätsfamilie)
Sei $R^{\perp}$ eine Menge von Rollen inkl. einem Don't-Care Symbol $\perp$, $G = (V_G, E_G)$ ein Graph, $M = (V_M, E_M)$ ein Motiv und $\mathcal{G}_M$ die Treffermenge zu G und M. Zusätzlich sei eine Abbildung role: $V_G \times \mathcal{G}_M \mapsto R^{\perp}$*, die jedem Knoten unter einem festen Treffer eine Rolle oder $\perp$ (falls der Knoten nicht im Treffer auftritt) zuweist, gegeben. Die* Rollen-basierte Motiv-basierte Zentralitätsfamilie basierend auf dem Motiv *M (kurz: Rollen-basierte Zentralität)*

(a) (b)

Abbildung 5.4.: (a) Das Feed-Forward Loop Motiv mit den Rollen A, B und C und (b) der Beispielgraph aus der Abbildung 5.3(b).

	Zentralitätswerte		
Knoten	Rolle A	Rolle B	Rolle C
v_1	1	0	0
v_2	2	1	0
v_3	0	1	1
v_4	0	0	2
v_5	0	1	0

Tabelle 5.2.: Die Zentralitätswerte für den Graphen aus Abbildung 5.4(b) berechnet mit der Rollen-basierten Zentralität für das Feed-Forward Loop Motiv mit Rollen (Siehe Abb. 5.4(a)).

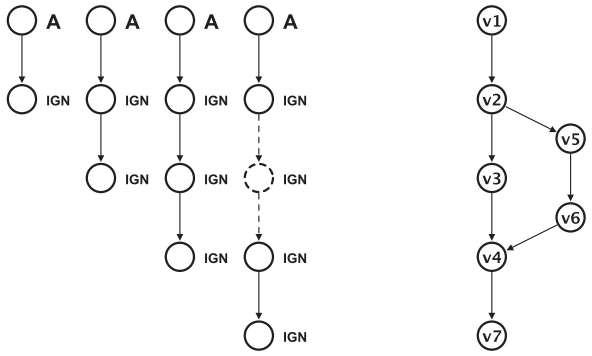
ist dann definiert als $\mathcal{C}^M_{rmc}(v,r)\colon (V_G \times R) \mapsto \mathbb{R}$ *mit* $\mathcal{C}^M_{rmc}(v,r) := |\{G_M \mid G_M \in \mathcal{G}_M \wedge v \in V(G_M) \wedge \text{role}(v, G_M) = r\}|$.

Die Rollen-basierte Zentralität zählt bei der Ermittlung der Zentralitätswerte somit die Anzahl der Treffer in der ein Knoten unter einer bestimmten Rolle vorkommt. Erst durch das Fixieren einer Rolle wird diese Zentralitätsfamilie zu einer Zentralität. Im Folgenden wird auch die Zentralitätsfamilie als Zentralität bezeichnet; die notwendige Fixierung einer Rolle wird dabei implizit angenommen.

Tabelle 5.2 zeigt die Zentralitätswerte der Rollen-basierten Zentralität berechnet für den Beispielgraphen aus der Abbildung 5.4(b) unter Berücksichtigung des FFL Motivs mit Rollen wie in der Abbildung 5.4(a) dargestellt. Die Komponenten der Matrix geben die Anzahl der Vorkommen der Knoten $v_1, \ldots, v_5$ in den Rollen A bis C an. Bezüglich der Rolle A, in der ein Knoten die beiden anderen Knoten kontrollieren kann, ist der Knoten v_2 der zentralste, gefolgt vom Knoten v_1. Diese „Fähigkeit zur Kontrolle der anderen Knoten“ ist bei der Betrachtung der Zentralitätswerte ohne Berücksichtigung der Rollen (Tabelle 5.1) nicht erkennbar.

5.2.3. Rollen-basierte Motiv-basierte Zentralitätsfamilie für Klassen von Motiven

Durch die Verwendung mehrerer „ähnlicher“ Motive wird das Konzept der Rollen-basierten Motiv-basierten Zentralitätsfamilie erweitert. Hierzu wird dieselbe Rolle an „ähnliche“ Knoten in einer ganzen Gruppe von „ähnlichen“ Motiven vergeben. Dadurch wird eine Zentralitätsfamilie über Gruppen von Motiven oder *Motiv-Klassen* definiert. In der Abbildung 5.5(a) ist eine solche Motiv-Klasse, die Motiv-Klasse *Ketten*, exemplarisch dargestellt. Bei dieser Klasse von Motiven wird jeweils der Startknoten (markiert mit der Rolle A) ausgezeichnet und für die Zentralitätsberechnung genutzt, da dieser Knoten (innerhalb seines Motivs betrachtet) der einzige ist, der alle anderen Knoten (innerhalb des Motivs) erreichen kann.



(a) (b)

Abbildung 5.5.: (a) Eine schematische Darstellung der ersten Motive der Motiv-Klasse *Ketten*. Bei dieser Motiv-Klasse sollen die mit der Rolle A bezeichneten Knoten für die Berechnung der Zentralität genutzt und die mit der Rolle IGN bezeichneten Knoten nicht berücksichtigt werden. (b) Ein Graph, um das Konzept der Motiv-Klassen-Zentralität für Ketten zu illustrieren.

Alle mit IGN markierten Knoten werden hingegen bei der Berechnung der Zentralitätswerte ignoriert.

Formal lässt sich eine Zentralitätsfamilie über Gruppen von „ähnlichen“ Motiven auf mindestens zwei Arten definieren: einerseits über die Festlegung einer Menge von Motiven und die Reduktion auf die Rollen-basierte Zentralität und andererseits über die Betrachtung einer durch eine Graphgrammatik [42] beschriebenen Sprache. Da die Formalisierung über Graphgrammatiken keine wesentlichen Vorteile bringt, wird im Folgenden auf diese verzichtet.

Definition 5.5 (Rollen-basierte Motiv-basierte Z.-familie für Motiv-Klassen) *Sei $R^{\perp}$ eine Menge von Rollen inkl. einem Don't-Care Symbol $\perp$, $G = (V_G, E_G)$ ein Graph, $\mathcal{M}$ eine Menge von Motiven und für alle $M \in \mathcal{M}$ sei $\mathcal{G}_M$ die Treffermenge zu G und M. Zusätzlich sei zu jedem Motiv $M \in \mathcal{M}$ eine Abbildung* $\text{role}^M \colon V_G \times \mathcal{G}_M \mapsto R^{\perp}$, *die jedem Knoten unter einem festen Treffer eine Rolle oder $\perp$ (falls der Knoten nicht im Treffer auftritt) zuweist, gegeben. Die* Rollen-basierte Motiv-basierte Zentralitätsfamilie basierend auf der Motivklasse $\mathcal{M}$ *(kurz: Motiv-Klassen Zentralität) ist dann definiert als* $\mathcal{C}^{\mathcal{M}}_{mcc}(v, r) \colon (V_G \times R) \mapsto \mathbb{R}$ *mit* $\mathcal{C}^{\mathcal{M}}_{mcc}(v, r) := \sum_{M \in \mathcal{M}} \mathcal{C}^{M}_{rmc}(v, r)$.

Im Graph in der Abbildung 5.5(b) sind Ketten der maximalen Länge sechs, d.h. mit sechs Knoten, möglich. Die Zentralitätswerte für die einzelnen Ketten und für die Motiv-Klassen Zentralität, jeweils für den Knoten am Anfang der Ketten (Rolle A), sind in der Tabelle 5.3 aufgeführt. Obwohl der Knoten v_1 nur einen Ausgangsgrad von 1 hat, wird dieser durch die Motiv-Klassen Zentralität für Ketten am höchsten bewertet, da dieser Knoten der einzige ist, von dem aus alle anderen Knoten erreicht werden können.

5.3. Algorithmen zur Berechnung der Motiv-basierten Zentralitäten

Aus den Definitionen der drei Zentralitäten lassen sich die notwendigen Algorithmen zur Berechnung der Zentralitätswerte ableiten. Hierzu werden zuerst Verfahren zur Berechnung der

		Kettenlänge				
Knoten	$\mathcal{C}_{mcc}^{\text{Ketten}}$	2	3	4	5	6
v_1	8	1	2	2	2	1
v_2	7	2	2	2	1	0
v_3	2	1	1	0	0	0
v_4	1	1	0	0	0	0
v_5	3	1	1	1	0	0
v_6	2	1	1	0	0	0
v_7	0	0	0	0	0	0

Tabelle 5.3.: Die Zentralitätswerte für den Graphen in Abbildung 5.5(b) berechnet mit der Motiv-Klassen Zentralität für Ketten unter Betrachtung der Rolle A (siehe Abb.5.5(a)).

Treffermenge und darauf aufbauend drei naive Algorithmen für die Berechnung der Zentralitätswerte vorgestellt. Für jeden Algorithmus wird jeweils die Laufzeit- und die Speicherplatzkomplexität angegeben. In Hinblick auf die Laufzeit- oder Speicherplatzkomplexität optimierte Algorithmen werden in diesem Abschnitt nicht vertiefend betrachtet.

5.3.1. Algorithmen zur Berechnung der Treffermenge

Die Berechnung der Treffermenge eines Motivs zu einem gegebenen Quellgraphen erfordert einen erheblichen Aufwand in Bezug auf die Laufzeit- und die Speicherplatzkomplexität, denn bereits das Entscheidungsproblem mit der Fragestellung „Existiert zu einem gegebenen Graphen in einem gegebenen Quellgraphen ein isomorpher Teilgraph“ ist ein NP-vollständiges Problem (Subgraph Isomorphism [30, 61]). Hieraus folgt unmittelbar, dass für die Berechnung aller Treffer ebenfalls kein effizienter Algorithmus angegeben werden kann.

Für den allgemeinen Fall, d.h. für einen nicht weiter eingeschränkten Quellgraphen mit $n := |V(G)|$ Knoten und einem Motiv mit $k := |V(M)|$ Knoten, liegt die obere Schranke der Laufzeit eines naiven Algorithmus für die Berechnung der Treffermenge bei $\mathcal{O}(n^k)$, da alle k Variationen ohne Wiederholung aus den Knotenmengen der beiden Graphen als mögliche Kandidaten auf Isomorphie überprüft werden müssen [126]. Etwas präziser lässt sich die Grenze mit $\mathcal{O}(\frac{n!}{(n-k)!})$ angeben, da in der Abschätzung durch $\mathcal{O}(n^k)$ auch doppelte Vorkommen („Variationen mit Wiederholungen“) erlaubt sind und diese für die Teilgraph-Isomorphie, bei der ein Isomorphismus zwischen zwei Knotenmengen mit k Knoten zu bestimmen ist, nicht zulässig sind. Es gilt allerdings $\mathcal{O}(\frac{n!}{(n-k)!}) \subseteq \mathcal{O}(n^k)$ und folglich ist die (gröbere) Abschätzung mit $\mathcal{O}(n^k)$ zulässig. Eine Abschätzung mit Methoden der parametrisierten Komplexitätstheorie führt (wahrscheinlich) zu keiner besseren Abschätzung, denn das Problem der Teilgraph-Isomorphie liegt in der Komplexitätsklasse $A[1]$. Diese Komplexitätsklasse ist identisch zur Klasse $W[1]$, der Klasse der Probleme die (wahrscheinlich) nicht *Fixed-Parameter Tractable* (FPT) sind [52].

Es existieren eine Reihe von Algorithmen, die für den allgemeinen Fall durch rechtzeitiges Erkennen von Knotenkombinationen, die nicht zu einem Isomorphismus zwischen dem zu suchenden Teilgraphen und dem Quellgraphen führen können, die entsprechende Teilberechnung vorzeitig beenden. Der bekannteste Algorithmus hierzu wurde 1976 von Ullmann vorgestellt [163]. Cordella *et al.* haben einen Algorithmus vorgestellt [32, 33], der im Vergleich zum Vorschlag von Ullmann ein verbessertes Laufzeitverhalten in vielen getesteten Beispielsituationen zeigt [31, 53, 54, 149]. Hardware-nahe Realisierungen von Algorithmen,

die den allgemeinen Fall des Problems lösen, wurden ebenfalls bereits untersucht und zeigen ein deutlich verbessertes Laufzeitverhalten gegenüber rein Software-basierten Implementationen [73]. Für alle diese Algorithmen gilt allerdings weiterhin, dass die Laufzeitkomplexität im Worst-Case $\mathcal{O}(n^k)$ beträgt.

Für ausgewählte Klassen von Graphen, beispielsweise wenn sowohl das Motiv als auch der Quellgraph Bäume sind oder wenn der Quellgraph planar ist, existieren spezialisierte Algorithmen für die Berechnung der Teilgraph-Isomorphie [44, 157]. Für die Analyse molekularbiologischer Netzwerke spielen diese Spezialalgorithmen allerdings nur eine untergeordnete Rolle, da im Allgemeinen nicht davon ausgegangen werden kann, dass ein zu untersuchendes Netzwerk die entsprechenden Bedingungen erfüllt. Es stehen somit im Allgemeinen nur die Algorithmen für nicht eingeschränkte Graphen zur Verfügung.

Die oben genannten Algorithmen von Ullmann bzw. Cordella *et al.* berechnen pro Aufruf je einen Isomorphismus zwischen zwei Knotenmengen. Erst durch Iteration, die von den Algorithmen unterstützt wird, d.h. Zwischenergebnisse bleiben erhalten und können wiederverwendet werden, ist es möglich, die für die Motiv-basierten Zentralitäten notwendige Treffermenge zu berechnen. Von Cortadella und Valiente wurde ein Algorithmus vorgeschlagen, der als zentrale Datenstruktur binäre Entscheidungsdiagramme (engl. *Binary Decision Diagrams, BDDs*) [27] verwendet und die vollständige Treffermenge ermittelt [34]. Hierbei werden die Knoten- und Kantenmengen und auch die Isomorphismen zwischen zwei Knotenmengen durch BDDs modelliert. Aufgrund der Speicherplatzeffizienz der BDDs bietet sich diese Art der Implementation insbesondere für den Fall eines kleinen Motivs und eines großen Quellgraphen an [34].

Wie bereits im Anschluss an die Definition der Treffermenge (Definition 5.2) erläutert, enthält diese nicht notwendigerweise alle Teilgraph-Isomorphismen zwischen dem Motiv und dem Quellgraphen, sondern nur die unterscheidbaren Teilgraphen. Die oben genannten Algorithmen berechnen allerdings immer alle möglichen Teilgraph-Isomorphismen. Unter Umständen ist deshalb eine Reduktion der gefundenen Teilgraph-Isomorphismen auf die unterscheidbaren Teilgraphen notwendig. Hierzu müssen entweder zuerst alle Teilgraph-Isomorphismen berechnet und danach eine Elimination der Dubletten durchgeführt werden oder es muss bereits während der iterativen Berechnung der Isomorphismen entschieden werden, welche Teilgraphen in die Treffermenge aufgenommen werden sollen. Beide Vorgehen erfordern dabei aber zwingend, dass alle Teilgraph-Isomorphismen berechnet werden.

Im Folgenden wird nun der Algorithmus zur Berechnung der Treffermenge, d.h. die Berechnung der Teilgraph-Isomorphismen und die anschließende Reduktion auf die unterscheidbaren Teilgraphen, als gegeben angesehen und mit getMatchSet bezeichnet. Dieser Algorithmus hat eine abgeschätzte Laufzeitkomplexität von $\mathcal{O}(n^k)$ und eine angenommene Speicherplatzkomplexität von $\mathcal{O}(k * n^k)$, da für die maximal n^k möglichen Teilgraph-Isomorphismen jeweils die Zuordnung der Knoten des Motivs zu den Knoten des Quellgraphens gespeichert werden müssen.

5.3.2. Naiver Algorithmus für die Motiv-basierte Zentralität

Unter Verwendung des oben beschriebenen Verfahrens zur Berechnung der Treffermenge $\mathcal{G}_M$ lassen sich die Zentralitätswerte der Knoten von G bzgl. dem Motiv M durch Algorithmus 5.1 berechnen. Für diesen Algorithmus kann die Laufzeitkomplexität für die Schritte 2–3 mit $\mathcal{O}(n)$, $\mathcal{O}(n^k)$ für die Berechnung der Treffermenge im Schritt 5 und $\mathcal{O}(k * n^k)$ für die Schritte 7–9 abgeschätzt werden. Die Speicherplatzkomplexität lässt sich mit $\mathcal{O}(n)$ für die Schritte 2–3 und $\mathcal{O}(k * n^k)$ für den Schritt 5 abschätzen. Daraus ergibt sich eine

Algorithmus 5.1 Motiv-basierte Zentralität

Eingabe: G: Graph, M: Motiv
Ausgabe: $\mathcal{C}_{mc}^{M}(v)$: Zentralitätswerte für die Knoten $v \in V(G)$

```
1: // Initialisiere den Resultatvektor
2: for all v ∈ V(G) do
3:     C_mc^M[v] ← 0
4: // Berechne die Treffermenge G_M
5: G_M ← getMatchSet(G,M)
6: // Berechne die Zentralitätswerte
7: for all G_M ∈ G_M do
8:     for all v ∈ V(G_M) do
9:         C_mc^M[v] ← C_mc^M[v] + 1
```

Algorithmus 5.2 Rollen-basierte Zentralität

Eingabe: G: Graph, M: Motiv mit Rollen R
Ausgabe: $\mathcal{C}_{rmc}^{M}(v, r)$: Zentralitätswerte für die Knoten $v \in V(G)$ und die Rollen $r \in R$

```
1: // Initialisiere die Resultatmatrix
2: for all v ∈ V(G) do
3:     for all r ∈ R do
4:         C_rmc^M[v, r] ← 0
5: // Berechne die Treffermenge G_M
6: G_M ← getMatchSet(G,M)
7: // Berechne die Zentralitätswerte
8: for all G_M ∈ G_M do
9:     for all v ∈ V(G_M) do
10:        r ← getRoleOfMatchingVertex(v, G_M)
11:        C_rmc^M[v, r] ← C_rmc^M[v, r] + 1
```

abgeschätzte Laufzeitkomplexität für den gesamten Algorithmus von $\mathcal{O}(k * n^k)$ und eine abgeschätzte Speicherplatzkomplexität von ebenfalls $\mathcal{O}(k * n^k)$.

5.3.3. Naiver Algorithmus für die Rollen-basierte Zentralität

Der Algorithmus für die Rollen-basierte Zentralität baut auf dem Algorithmus für die Motiv-basierte Zentralität auf (Algorithmus 5.1) und ist in Algorithmus 5.2 dargestellt. Die Funktion getRoleOfMatchingVertex in Schritt 10 berechnet zu einem gegebenen Knoten $v \in V(G)$ und dem Treffer G_M die entsprechende Rolle, die der Knoten innerhalb des Treffers einnimmt. Sie entspricht dabei genau der Funktion *role* in der Definition 5.4. Für diese Funktion kann eine Laufzeit- und eine Speicherplatzkomplexität von $\mathcal{O}(1)$ angenommen werden.

Im Unterschied zum Algorithmus 5.1 liefert dieser Algorithmus eine Matrix mit Zentralitätswerten zurück. Die Zeilen dieser Matrix repräsentieren die Knoten des Graphen G, die Spalten die Rollen des Motivs und in den einzelnen Einträgen stehen die berechneten Zentralitätswerte. Die Laufzeitkomplexität dieses Algorithmus liegt in derselben Klasse wie die von Algorithmus 5.1, da die Laufzeitkomplexität der Schritte 2–4 mit $\mathcal{O}(n*|R|)$ abgeschätzt werden kann[3], die Funktion getRoleOfMatchingVertex (Schritt 10) als $\mathcal{O}(1)$-Operation angenommen werden kann und alle anderen Schritte gegenüber Algorithmus 5.1 unverän-

[3] Die Anzahl der möglichen Rollen für die Knoten eines Motivs ist durch die Anzahl der Knoten beschränkt, d.h. es gilt immer $|R| \leq k = |V(M)|$.

Algorithmus 5.3 Motiv-Klassen Zentralität

Eingabe: G: Graph, $\mathcal{M}$ Menge von Motiven mit Rollen aus R
Ausgabe: $\mathcal{C}^{\mathcal{M}}_{mcc}(v,r)$: Zentralitätswerte für die Knoten $v \in V(G)$ und die Rollen $r \in R$

```
1: // Initialisiere die Resultatmatrix
2: for all v ∈ V(G) do
3:     for all r ∈ R do
4:         C^M_mcc[v, r] ← 0
5: // Berechne die Zentralitätswerte
6: for all M ∈ M do
7:     tempCent ← computeRoleMotifCent(G, M) // Algo. 5.2
8:     for all r ∈ R do
9:         for all v ∈ V(G) do
10:            C^M_mcc[v, r] ← C^M_mcc[v, r] + tempCent[v, r]
```

dert sind. Die Speicherplatzkomplexität unterscheidet sich ebenfalls nicht wesentlich von Algorithmus 5.1, da die Matrix zur Speicherung des Resultats einen Speicherplatzbedarf von $\mathcal{O}(n * |R|)$ hat und alle anderen Schritte sich hinsichtlich der Speicherplatzkomplexität nicht unterscheiden. Folglich hat dieser Algorithmus ebenfalls eine angenommene Worst-Case Laufzeitkomplexität von $\mathcal{O}(k * n^k)$ und eine angenommene Worst-Case Speicherplatzkomplexität von $\mathcal{O}(k * n^k)$.

5.3.4. Naiver Algorithmus für die Motiv-Klassen Zentralität

Die Motiv-Klassen Zentralität wird mit Hilfe von Algorithmus 5.3 berechnet. Wie schon der Algorithmus für die Rollen-basierte Zentralität liefert auch dieser Algorithmus eine Matrix mit Zentralitätswerten zurück.

Unter Zuhilfenahme der Definition $k_{\max} := \max\{|V(M)| \mid M \in \mathcal{M}\}$, d. h. $k_{\max}$ entspricht der maximalen Knotenanzahl aller Motive in der Menge der gewählten Motive, lässt sich die Laufzeitkomplexität des Algorithmus mit $\mathcal{O}(k_{\max} * n^{k_{\max}})$ abschätzen. Dies folgt unmittelbar aus der Abschätzung der Laufzeitkomplexität von Algorithmus 5.2, denn fürs jedes Motiv $(i = 1, \ldots, l)$ muss für die Berechnung der Motiv-Klassen-Zentralität Algorithmus 5.2 aufgerufen werden. Dieser Algorithmus hat eine Laufzeitkomplexität von $\mathcal{O}(k_i * n^{k_i})$. Folglich lässt sich die Gesamtlaufzeitkomplexität mit $\mathcal{O}(k_{\max} * n^{k_{\max}}) = \sum_{i=1}^{l} \mathcal{O}(k_i * n^{k_i})$ abschätzen. Die Speicherplatzkomplexität kann analog zur Speicherplatzkomplexität von Algorithmus 5.2 ebenfalls mit $\mathcal{O}(k_{\max} * n^{k_{\max}})$ abgeschätzt werden.

5.3.5. Optimierte Algorithmen für die Berechnung der Zentralitäten

Die vorgestellten naiven Algorithmen zur Berechnung der Zentralitätswerte können sicherlich noch verbessert werden. Zwei mögliche Richtungen der Optimierung sind a. die unmittelbare Auswertung der gefundenen Treffer während der Berechnung der Teilgraph-Isomorphismen und b. die Verwendung von Algorithmen, die für ausgewählte Motive optimiert sind.

Im ersten Fall, d.h. bei der Verwendung von Algorithmen, die die einzelnen Treffer nicht speichern, sondern unmittelbar auswerten, könnte insbesondere die aufwändige Speicherung und nachträgliche Reduktion der Treffermenge entfallen. Ein entsprechender Algorithmus müsste hierzu allerdings die unterscheidbaren Teilgraphen schon bei der Berechnung der Treffer erkennen bzw. den dadurch entstehenden Effekt der mehrfachen Berücksichtigung der Treffer nachträglich wieder „herausrechnen“. Umgesetzt werden könnte dieses über die

Betrachtung der zum Motiv gehörenden Automorphismen-Gruppe. Wenn bereits vor der Berechnung der Zentralitätswerte die automorphen Motiv-Knoten bekannt wären, dann könnten die mehrfach gezählten Treffer nachträglich abgezogen werden.

Der zweite Fall, d.h. die Verwendung von optimierten Algorithmen für das Auffinden von Treffern für bestimmte Motive, wurde im Einzelfall bereits betrachtet. Zwei Beispiele sind die von Alon *et al.* veröffentlichten Algorithmen und Schranken für das Finden und Zählen von Zyklen in gerichteten und ungerichteten Graphen [8] und die von Chiba und Nishizeki veröffentlichten Algorithmen zum Auffinden von Dreiecken, Quadraten und vollständigen Graphen [28].

5.4. Analyse eines Genregulationsnetzwerkes von *E. coli* mittels Motiv-basierten Zentralitäten

Escherichia coli (*E. coli*) gehört mit zu den am besten untersuchten Organismen. Aus den für diesen Organismus bekannten Informationen über die Genregulation lässt sich ein Genregulationsnetzwerk (siehe Abschnitt 3.4.1), d.h. ein Netzwerk, das die gegenseitige Beeinflussung von Genen bzw. deren Produkten im Rahmen der Genexpression beschreibt, ableiten. Im Folgenden wird ein solches Netzwerk beschrieben und mit Hilfe der vorgestellten Motiv-basierten Zentralitäten untersucht. Durch die Analyse werden Informationen zu wichtigen genregulatorischen Vorgängen für *E. coli* aufgedeckt. Abschließend werden die mithilfe der vorgestellten Zentralitätsmaße ermittelten Resultate noch mit Resultaten, die unter Verwendung von bereits bekannten Zentralitätsmaßen gewonnen wurden, verglichen.

5.4.1. Das Genregulationsnetzwerk für *E. coli*, globale Regulatoren und Motive für dieses Netzwerk

Das Genregulationsnetzwerk für *E. coli* beschreibt die während der Transkription auftretenden regulatorischen Interaktionen zwischen Genen. Für den Organismus *E. coli* sind viele der bekannten Interaktionen in der Datenbank RegulonDB [147] dokumentiert und dort als Interaktionspaare zwischen Transkriptionsfaktoren und Genen[4] abgelegt. Aus diesen Informationen wird ein Netzwerk erstellt, in dem jeder Knoten ein Gen repräsentiert. Alle Kanten in diesem Netzwerk sind gerichtet und repräsentieren die Kontrolle, die ein Transkriptionsfaktor, der Startknoten einer Kante, über das kontrollierte Gen, den Endknoten der Kante, hat. Aus den einzelnen Interaktionspaaren wird ein gerichteter Graph, da Gene, die als Endpunkte einer Kante auftreten, selbst wieder Transkriptionsfaktoren kodieren können. Ein Spezialfall bei der Aufbereitung sind zusammengesetzte Transkriptionsfaktoren. Diese bestehen aus mehreren Untereinheiten, die selbst Produkte unterschiedlicher Gene sind. Zusammengesetzten Transkriptionsfaktoren werden in der Modellierung als ein Transkriptionsfaktor dargestellt; der zusammengesetzte Transkriptionsfaktor aus den beiden Produkten der Gene *ihfA* und *ihfB* wird beispielsweise als Transkriptionsfaktor *ihfAB* im Netzwerk repräsentiert. Alle Interaktionen, die zu Untereinheiten des Transkriptionsfaktors dokumentiert sind, werden zusammengefasst und dem zusammengesetzten Transkriptionsfaktor zugewiesen. Eventuell auftretende parallele Kanten werden danach eliminiert, da diese bei der Berechnung der Motiv-basierten Zentralitäten nicht von Bedeutung sind. Eine eventuell auftretende Selbstregulation eines Transkriptionsfaktors wird als Schleife im Netzwerk

[4] Im Folgenden werden der Einfachheit halber die Begriffe Gen und Transkriptionsfaktor synonym verwendet.

arcA, *cpxR*, *crp*, *cspA*, *fis*, *fnr*, *fur*, *ihfAB*,
hns, *lrp*, *mlc*, *narL*, *ompR*, *phoB*, *purR rob*, *soxR*, *soxS*,

Tabelle 5.4.: Von Martínez-Antonio & Collado-Vides publizierte Liste der globalen Regulatoren für *E. coli*. Quelle: [120]

abgebildet. Während der Berechnung der Zentralitätswerte auf der Basis des FFL Motivs und der regulatorischen Ketten werden diese Schleifen allerdings nicht berücksichtigt, da sie nicht Bestandteil der Motive sind. Das aufbereitete Netzwerk besteht aus insgesamt 1250 Knoten und 2515 Kanten.

Gene, die einen weitreichenden Einfluss in der Genregulation haben, werden *globale Regulatoren* genannt. Einige Methoden um diese für einen Organismus zu bestimmen sind in der Literatur [120] benannt. Hierzu gehören beispielsweise die Betrachtung der Anzahl der regulierten Gene, die Anzahl und die Typen der genutzten Ko-Regulatoren, die Anzahl der kontrollierten Regulatoren, die Größe der evolutionären Familie, zu der ein Gen gehört, und die Vielfältigkeit der Umweltbedingungen, unter denen das betrachtete Gen aktiv ist. Für *E. coli* haben Martínez-Antonio & Collado-Vides anhand dieser Kriterien insgesamt 18 globale Regulatoren durch Auswertung verschiedener Datenquellen ermittelt und beschrieben [120]. Diese sind in der Tabelle 5.4 aufgeführt.

Innerhalb von Genregulationsnetzwerken von *E. coli* und *Saccharomyces cerevisiae* (*S. cerevisiae*, „Bäckerhefe“) tritt das FFL Motiv häufiger auf, als statistisch zu erwarten ist [103, 158]. In diesem Motiv (siehe Abbildung 5.4(a)) nehmen die drei Knoten drei unterschiedliche Rollen an. Der oben stehende Knoten (A) reguliert die beiden anderen Knoten und ist somit der Hauptregulator (engl. *master regulator*). Der Knoten auf der rechten Seite (B) wird vom Hauptregulator reguliert und reguliert zusammen mit diesem den dritten Knoten. Der unten dargestellte Knoten (C) wird durch die beiden anderen Knoten reguliert. Je nach Typ der Interaktion zwischen diesen drei Knoten, d.h. Aktivierung oder Unterdrückung der Expression des kontrollierten Gens, kann dieses Motiv als Beschleuniger oder als Verzögerer auf die Expression des dritten Gens wirken [116, 117].

Zusätzlich zu Motiven mit einer festen Anzahl an Knoten wurden auch Motive mit einer variablen Anzahl von Knoten, beispielsweise die Motive *single input module* (SIM) und *regulatory chain* (siehe Abschnitt 5.2.3), in Genregulationsnetzwerken als relevant identifiziert [103, 158]. Beide Motive können als Motiv-Klassen modelliert und zur Berechnung der Motiv-Klassen Zentralität eingesetzt werden.

5.4.2. Motiv-basierte Zentralität für das *E. coli* Genregulationsnetzwerk

Die Namen der Top 2% der Gene des *E. coli* Genregulationsnetzwerkes auf der Basis der Motiv-basierten Zentralität für das FFL Motiv $\mathcal{C}_{mc}^{\mathrm{FFL}}$ sind in der Tabelle 5.5 dargestellt. Der Vergleich der Namen mit der Liste der globalen Regulatoren (Tabelle 5.4 und 5.7) zeigt, dass durch die Verwendung der Motiv-basierten Zentralität für das FFL Motiv bereits eine Reihe von globalen Regulatoren identifiziert werden können.

5.4.3. Rollen-basierte Zentralität für das *E. coli* Genregulationsnetzwerk

Die Rollen-basierte Zentralität $\mathcal{C}_{rmc}^{\mathrm{FFL}}$ für das FFL Motiv mit Rollen (siehe Abbildung 5.4(a)) liefert für jedes Gen drei Zentralitätswerte, je einen Wert für die Rollen A, B und C. Aus diesen Zentralitätswerten ergeben sich folglich drei unterschiedliche Reihenfolgen der Gene,

	$\mathcal{C}_{mc}^{\text{FFL}}$	$\mathcal{C}_{rmc}^{\text{FFL(A)}}$	$\mathcal{C}_{rmc}^{\text{FFL(B)}}$	$\mathcal{C}_{rmc}^{\text{FFL(C)}}$	$\mathcal{C}_{mcc}^{\text{Ketten}}$
1	*crp*	*crp*	*narL*	*marB*	*crp*
2	*fnr*	*fnr*	*fis*	*gadA*	*ihfAB*
3	*arcA*	*ihfAB*	*arcA*	*fumB*	*arcA*
4	*fis*	*arcA*	*fnr*	*gadB*	*fnr*
5	*narL*	*fis*	*hns*	*gadC*	*fis*
6	*ihfAB*	*modE*	*fur*	*lpdA*	*evgA*
7	*hns*	*soxS*	*hyfR*	*sodA*	*ydeO*
8	*fur*	*hns*	*gadX*	*aceE*	*gadE*
9	*gadX*	*cpxR*	*marA*	*aceF*	*soxR*
10	*hyfR*	*fhlA*	*tdcA*	*flhC*	*soxS*
11	*marA*	*gadE*	*yiaJ*	*flhD*	*torR*
12	*flhD*	*rob*	*flhD*	*glpA*	*gadW*
13	*nagC*	*gadX*	*galS*	*glpB*	*cspE*
14	*soxS*	*galR*	*nagC*	*glpC*	*cspA*
15	*modE*	*fur*	*idnR*	*gutQ*	*gadX*
16	*tdcA*	*gntR*	*ompR*	*hdeA*	*hns*
17	*yiaJ*	*oxyR*	*cytR*	*hdeB*	*oxyR*
18	*gutM*	*tdcR*	*gutM*	*mglA*	*fur*
19	*ompR*	*gutM*	*srlR*	*mglC*	*modE*
20	*srlR*	*nagC*	*caiF*	*mtlA*	*narL*
21	*galS*	*narL*	*chbR*	*mtlD*	*lrp*
22	*idnR*	*ompR*	*nikR*	*nuoA*	*glnG*
23	*caiF*	*srlR*	*glpR*	*nuoB*	*ompR*
24	*chbR*	*argP*	*malT*	*nuoC*	*phoB*
25	*cpxR*	*cysB*	*araC*	*nuoE*	*cpxR*

Tabelle 5.5.: Namen der Top 2% der Gene von *E. coli* bzgl. der Motiv-basierten Zentralität auf der Basis des Feed-Forward Loop Motivs ($\mathcal{C}_{mc}^{\text{FFL}}$), der Rollen-basierten Zentralität für das Feed-Forward Loop Motiv mit Rollen ($\mathcal{C}_{rmc}^{\text{FFL(A)}}$, $\mathcal{C}_{rmc}^{\text{FFL(B)}}$ und $\mathcal{C}_{rmc}^{\text{FFL(C)}}$) und der Motiv-Klassen Zentralität für Ketten ($\mathcal{C}_{mcc}^{\text{Ketten}}$). In den Fällen, in denen Gene identische Zentralitätswerte zugewiesen wurden, erfolgt die Auflistung in alphabetischer Reihenfolge.

siehe Tabelle 5.5.

Die Gene an den oberen Positionen der Rollen A und B können dabei in drei Gruppen aufgeteilt werden: einige Gene (*crp*, *ihfAB*, *soxS*) nehmen fast ausschließlich die Rolle A ein und sind somit eher Hauptregulatoren im Regulationsnetz von *E. coli*; sie werden nicht bzw. kaum durch andere Gene reguliert. Einige Gene (*gadX*, *fur*, *hyfR*, *narL*) nehmen fast ausschließlich die Rolle B ein und regulieren andere Gene nur in Verbindung mit einem zweiten Gen. Die Gene der dritten Gruppe (*arcA*, *fis*, *fnr*, *hns*) nehmen beide Rollen A und B ein. Wahlweise treten diese Gene also als Hauptregulatoren auf oder sie regulieren in Verbindung mit einem anderen Regulator.

Unter Zuhilfenahme der Rollen des FFL Motivs ist somit eine genauere Analyse der Regulationsweise der Gene innerhalb des Regulationsnetzes möglich. Insbesondere die Unterscheidung in Hauptregulatoren und Genen, die nur in Verbindung mit anderen Genen wirken, ist hierdurch möglich.

		Länge der Kette					
Gen	$\mathcal{C}_{mcc}^{\text{Ketten}}$	2	3	4	5	6	7
crp	1592	359	525	436	212	60	0
ihfAB	667	186	215	156	82	28	0
arcA	470	111	215	127	17	0	0
fnr	470	206	237	27	0	0	0
fis	387	156	121	82	28	0	0
evgA	325	4	27	90	125	51	28
ydeO	322	1	27	90	125	51	28
gadE	321	27	90	125	51	28	0
soxR	213	2	24	92	91	4	0
soxS	211	24	92	91	4	0	0
torR	191	10	15	87	51	28	0
gadW	185	4	15	87	51	28	0
cspE	184	1	2	88	65	28	0
cspA	183	2	88	65	28	0	0
gadX	181	15	87	51	28	0	0
hns	181	88	65	28	0	0	0
oxyR	166	15	73	74	4	0	0
fur	151	73	74	4	0	0	0
modE	141	32	94	15	0	0	0
narL	109	94	15	0	0	0	0
lrp	72	62	10	0	0	0	0
glnG	59	43	15	1	0	0	0
ompR	51	16	35	0	0	0	0
phoB	48	34	5	9	0	0	0
cpxR	45	34	11	0	0	0	0

Tabelle 5.6.: Namen und Zentralitätswerte der Top 2% der Gene des *E. coli* Genregulationsnetzes auf der Basis der Motiv-Klassen Zentralität auf der Basis der Motiv-Klasse Ketten für die Rolle *A* (siehe Abbildung 5.5(a)). Ketten länger als sieben Knoten treten im betrachteten Netzwerk nicht auf. Die Anzahl der Ketten der Länge 2 entspricht der Anzahl der Gene, die unmittelbar durch das betrachtete Gen reguliert wird. Dieser Wert ist identisch mit dem *Out-Degree*.

5.4.4. Motiv-Klassen Zentralität für das *E. coli* Genregulationsnetzwerk

Die Anwendung der Motiv-Klassen Zentralität auf der Basis der Motiv-Klasse *Ketten* auf das Genregulationsnetzwerk von *E. coli* ermöglicht die Ermittlung der Gene, die den größten Einfluss auf andere Gene haben. Die Zentralitätswerte bzgl. dieser Zentralität sind in der Tabelle 5.6 dargestellt.

Die Zusammensetzung der Zentralitätswerte aus den Werten der einzelnen Ketten zeigt dabei interessante Eigenschaften auf: Einige der Gene unter den Top 2% haben einen sehr geringen Zentralitätswert für die Ketten der Länge 2 (dieser entspricht genau dem *Out-Degree*). Hierzu gehören die Gene *cspA*, *cspE*, *evgA*, *gadW*, *soxR* und *ydeO*. Diese Gene haben einen geringen direkten Einfluss; folglich ist ihr Zentralitätswert auf der Basis des FFL Motivs ebenfalls gering. Allerdings haben alle genannten Gene eine hohe indirekte Kontrolle über andere Gene. Beispielsweise haben die Gene *evgA* und *ydeO*, Platz sechs und sieben in der Motiv-Klassen Zentralität für Ketten, einen *Out-Degree* von vier bzw. eins. Beide Gene regulieren allerdings das Gen *gadE* und sind auf diese Weise Teil derselben regulatorischen Kette. Zusätzlich reguliert das Gen *evgA* drei weitere Gene die selbst keine

weiteren Gene regulieren. Obwohl also die beiden Gene *evgA* und *ydeO* einen geringen *Out-Degree* haben, ist ihr indirekter Einfluss hoch. Die Motiv-Klassen Zentralität für Ketten ist somit in der Lage Gene zu bestimmen, die durch die *Out-Degree* Zentralität und die Motiv-basierte Zentralität für das FFL Motiv nicht als interessant gefunden wurden.

5.4.5. Vergleich der Bewertungen auf der Basis der Motiv-basierten Zentralitäten mit Bewertungen auf der Basis anderer Zentralitätsmaße

Da Genregulationsnetzwerke durch gerichtete Graphen modelliert werden und nicht notwendigerweise stark zusammenhängend sein müssen (siehe Abschnitt 3.4.1) können zusätzlich zu den Motiv-basierten Zentralitäten noch die Zentralitäten *In/Out-Degree*, *Katz Status Index*, *PageRank*, *Integration*, *Radiality* und *Shortest-Path Betweenness* zur Analyse dieser Netzwerke eingesetzt werden. Für die beiden Zentralitäten *Katz Status Index* und *PageRank* können dabei jeweils das gegebene Netzwerk und der dazugehörige inverse Graph (siehe Definition 2.15) betrachtet werden.

Die Namen der Top 25 Gene[5] unter Betrachtung der acht „besten“ Zentralitätsmaße[6] sind in der Tabelle 5.7 dargestellt.

Insgesamt wurden von Martínez-Antonio & Collado-Vides 18 globale Regulatoren beschrieben (siehe Tabelle 5.4). Von den betrachteten Zentralitätsmaßen aus Tabelle 5.7 sind alle in der Lage mehr als 50% der globalen Regulatoren unter die Top 2% der Gene zu platzieren. Die Zentralität *Shortest-Path Betweenness* identifiziert beispielsweise 11 globale Regulatoren und die Motiv-Klassen Zentralität auf der Basis der Motiv-Klasse Ketten ist in der Lage 15 der 18 globalen Regulatoren unter die Top 25 zu platzieren. Nahezu unter allen Zentralitätsmaßen werden die Top 5 Positionen durch globale Regulatoren besetzt, dennoch bewerten die einzelnen Zentralitätsmaße die Gene unterschiedlich. Beispielsweise wird das Gen *ihfAB* durch einige Zentralitäten (*PageRank*, *Radiality*) auf die zweite Position und durch andere Zentralitätsmaße (*Shortest-Path Betweenness*) noch nicht einmal unter die Top 2% platziert. Die Zentralität *Radiality* bewertet die Gene ähnlich zur Motiv-Klassen Zentralität für Ketten aber auch hier sind unterschiede offenkundig: der globale Regulator *fur* wird durch die *Radiality* auf Position 8 platziert, wohingegen die Motiv-Klassen Zentralität für Ketten dieses Gen auf Position 18 platziert.

Es lässt sich folglich feststellen, dass die betrachteten Zentralitäten die Gene unterschiedlich bewerten und das die Motiv-Klassen Zentralität für Ketten von allen betrachteten Zentralitätsmaßen die höchste Anzahl an globalen Regulatoren unter die Top 2% der betrachteten Gene platziert (15 von 18). Die Motiv-basierten Zentralitäten, insbesondere für die Motiv-Klasse Ketten, sind folglich besser als die bereits bekannten Zentralitätsmaße für die Analyse von Genregulationsnetzwerken geeignet.

5.5. Zusammenfassung

In diesem Kapitel wurden drei neue Zentralitätsmaße vorgestellt. Alle drei Maße sind durch die Wahl des Motivs „konfigurierbar“. Im Unterschied zu den existierenden Zentralitätsmaßen berücksichtigen die Motiv-basierten Zentralitäten jeweils nur einen Teilbereich des Netzes, sie füllen dadurch im Spektrum der Zentralitätsmaße eine noch bestehende Lücke

[5] Top 2% aller Gene aus dem betrachteten Netzwerk

[6] Zentralitätsmaße, die die größte Anzahl an globalen Regulatoren innerhalb der Top 2% aller Gene identifizieren

aus. Bestehende Zentralitäten betrachten nämlich entweder nur die unmittelbare Umgebung eines Knotens oder das gesamte Netzwerk. Im Vergleich dazu betrachten die Motiv-basierten Zentralitäten eine „konfigurierbare“ Umgebung um einen zu analysierenden Knoten.

Aufgrund der Tatsache, dass Motive innerhalb der Genregulation eine wichtige Rolle einnehmen, eignen sich die hier vorgestellten Zentralitätsmaße sehr gut zur Analyse von Genregulationsnetzwerken, wie am Beispiel des Netzwerkes für *E. coli* gezeigt werden konnte.

	$\mathcal{C}_{odeg}$	$\mathcal{C}_{rad}$	$\mathcal{C}_{spb}$	$\mathcal{C}_{katz}$	$\mathcal{C}_{pra}$	$\mathcal{C}_{mc}^{\mathrm{FFL}}$	$\mathcal{C}_{rmc}^{\mathrm{FFL(A)}}$	$\mathcal{C}_{mcc}^{\mathrm{Ketten}}$
1	***crp***	***crp***	***hns***	***crp***	***crp***	***crp***	***crp***	***crp***
2	***fnr***	***ihfAB***	*gadX*	***fnr***	***ihfAB***	***fnr***	***fnr***	***ihfAB***
3	***ihfAB***	***fnr***	*flhD*	***arcA***	***fnr***	***arcA***	***ihfAB***	***arcA***
4	***fis***	***arcA***	***fur***	***ihfAB***	***arcA***	***fis***	***arcA***	***fnr***
5	***arcA***	***fis***	*gadE*	***fis***	***phoB***	***narL***	***fis***	***fis***
6	***narL***	*gadE*	***fis***	***hns***	*lexA*	***ihfAB***	*modE*	*evgA*
7	***hns***	***hns***	***lrp***	*gadE*	***cpxR***	***hns***	***soxS***	*ydeO*
8	***fur***	***fur***	*rcsAB*	*gadX*	***soxR***	***fur***	***hns***	*gadE*
9	***lrp***	***soxS***	***soxS***	***cspA***	***fis***	*gadX*	***cpxR***	***soxR***
10	*glnG*	*evgA*	***fnr***	*evgA*	*evgA*	*hyfR*	*fhlA*	***soxS***
11	*narP*	*ydeO*	***cspA***	*ydeO*	*cysB*	*marA*	*gadE*	*torR*
12	***cpxR***	*oxyR*	*caiF*	*torR*	*argR*	*flhD*	***rob***	*gadW*
13	***phoB***	*gadX*	***purR***	*gadW*	*phoP*	*nagC*	*gadX*	*cspE*
14	*fruR*	***cspA***	***narL***	*cspE*	***fur***	***soxS***	*galR*	***cspA***
15	*modE*	***narL***	*marA*	***soxS***	*allR*	*modE*	***fur***	*gadX*
16	*fhlA*	*modE*	*metJ*	***soxR***	*glnG*	*tdcA*	*gntR*	***hns***
17	*lexA*	***soxR***	*malT*	***rob***	*sdaR*	*yiaJ*	*oxyR*	*oxyR*
18	*flhD*	*torR*	***arcA***	*marA*	*trpR*	*gutM*	*tdcR*	***fur***
19	*gadE*	*gadW*	*glnG*	*marR*	*agaR*	***ompR***	*gutM*	*modE*
20	***purR***	*cspE*	***ompR***	*oxyR*	*gadE*	*srlR*	*nagC*	***narL***
21	***soxS***	***lrp***	*nac*	***fur***	***soxS***	*galS*	***narL***	***lrp***
22	*argR*	*glnG*	*oxyR*	*modE*	***hns***	*idnR*	***ompR***	*glnG*
23	*cysB*	***phoB***	*hupAB*	*gutM*	***lrp***	*caiF*	*srlR*	***ompR***
24	*marA*	*narP*	*argP*	*srlR*	*tyrR*	*chbR*	*argP*	***phoB***
25	*nagC*	***ompR***	*dnaA*	***narL***	*torR*	***cpxR***	*cysB*	***cpxR***
# Glob. Reg.	13	14	11	12	12	11	12	15

Tabelle 5.7.: Namen der Top 2% der Gene von *E. coli* bzgl. der Zentralitäten *Out-Degree* ($\mathcal{C}_{odeg}$), *Radiality* ($\mathcal{C}_{rad}$), *Shortest-Path Betweenness* ($\mathcal{C}_{spb}$), *Katz Status Index* ($\mathcal{C}_{katz}$), *PageRank* ($\mathcal{C}_{pra}$), der Motiv-basierten Zentralität auf der Basis des Feed-Forward Loop Motivs ($\mathcal{C}_{mc}^{\mathrm{FFL}}$), der Rollen-basierten Zentralität für das Feed-Forward Loop Motiv mit der Rolle *A* ($\mathcal{C}_{rmc}^{\mathrm{FFL(A)}}$) und der Motiv-Klassen Zentralität für Ketten ($\mathcal{C}_{mcc}^{\mathrm{Ketten}}$). Bei den Zentralitäten *Katz Status Index* und *PageRank* wurde jeweils der inverse Graph (siehe Definition 2.15) verwendet. Globale Regulatoren nach Martínez-Antonio & Collado-Vides (siehe Tabelle 5.4) sind durch Fettdruck gekennzeichnet. In der letzten Zeile ist die durch die betreffende Zentralität identifizierte Anzahl an globalen Regulatoren (unter den Top 2% der Gene) angegeben. In den Fällen, in denen Genen identische Zentralitätswerte zugewiesen wurden, erfolgt die Auflistung in alphabetischer Reihenfolge. Quelle: [98] (Eigene Publikation)

6. Fluss-basierte Zentralitäten für die Analyse von metabolischen Reaktionsnetzen

Die Betrachtung einer Reihenfolge von Metaboliten oder Reaktionen eines metabolischen Reaktionsnetzes hilft, das Verständnis der Stoffwechselvorgänge eines Organismus zu verbessern. Im Folgenden wird deshalb ein neues Zentralitätsmaß zur Bestimmung einer Reihenfolge von Metaboliten eines gegebenen metabolischen Reaktionsnetzes vorgestellt. Von den bereits bekannten Zentralitätsmaßen unterscheidet sich diese Zentralität dahingehend, dass zusätzlich zur Netzstruktur Informationen über den Kohlenstoffffluss in die Berechnung der Reihenfolge mit einfließt. Hierbei handelt es sich allerdings nicht nur um eine einfache Erweiterung des Reaktionsnetzes um Kantengewichte. Um die biochemischen Prozesse innerhalb des Organismus korrekt abzubilden, wird aus dem zum Reaktionsnetz gehörenden bipartiten Graphen vor der Berechnung der Zentralitätswerte der zugehörige, den Kohlenstoffffluss beschreibende, gewichtete, unipartite Graph erstellt. Auf Basis dieses im Folgenden Metabolitgraphen mit Kohlenstoffffluss genannten Graphen werden dann die Zentralitätswerte für die Metaboliten ermittelt.

Die für die Berechnung der Zentralitätswerte notwendigen einzelnen Aufbereitungsschritte sind in der Abbildung 6.1 dargestellt. Ovale Knoten bezeichnen in dieser Abbildung Daten, beispielsweise Graphen oder Informationen über biochemische Reaktionen, und rechteckige Knoten bezeichnen Prozesse, d.h. Umwandlungen der gegebenen Daten. Die Beschreibung in diesem Kapitel folgt der Reihenfolge der Prozesse wie sie im Diagramm dargestellt sind.

Dieses Kapitel ist wie folgt strukturiert: Zuerst wird das Konzept des metabolischen Flusses erläutert (Abschnitt 6.1) und der Metabolitgraph mit Kohlenstoffffluss eingeführt (Abschnitt 6.2). Darauf aufbauend wird eine Zentralität zur Erstellung einer Reihenfolge von Metaboliten definiert (Abschnitt 6.3) und es werden die zur Berechnung notwendigen Algorithmen vorgestellt (Abschnitt 6.4). Abschließend wird die Zentralität zur Analyse eines Reaktionsnetzes von *E. coli* eingesetzt und die Resultate der Analyse mithilfe des neuen Zentralitätsmaßes werden mit Resultaten basierend auf den bereits bekannten Zentralitäten verglichen (Abschnitt 6.5).

6.1. Flüsse in metabolischen Reaktionsnetzen

Die Geschwindigkeit mit der eine biochemische Reaktion abläuft wird *metabolischer Fluss* (oder kurz *Fluss*) genannt. Gemessen werden kann dieser Fluss beispielsweise durch das Verfolgen (engl. *tracen*) von speziell in eine Zelle eingebrachten Metaboliten. Diese Metaboliten sind mit stabilen Isotopen, häufig ^{13}C, markiert und können mittels spezifischer Messmethoden (beispielsweise Massenspektrometrie oder kernmagnetischer Resonanz) beobachtet werden [150]. Abhängig ist der metabolische Fluss von einer Reihe von Bedingungen, beispielsweise den Stoffkonzentrationen der beteiligten Metaboliten, dem Vorhandensein einer entsprechenden Menge des katalysierenden Enzyms, regulatorischen Einflüssen durch andere

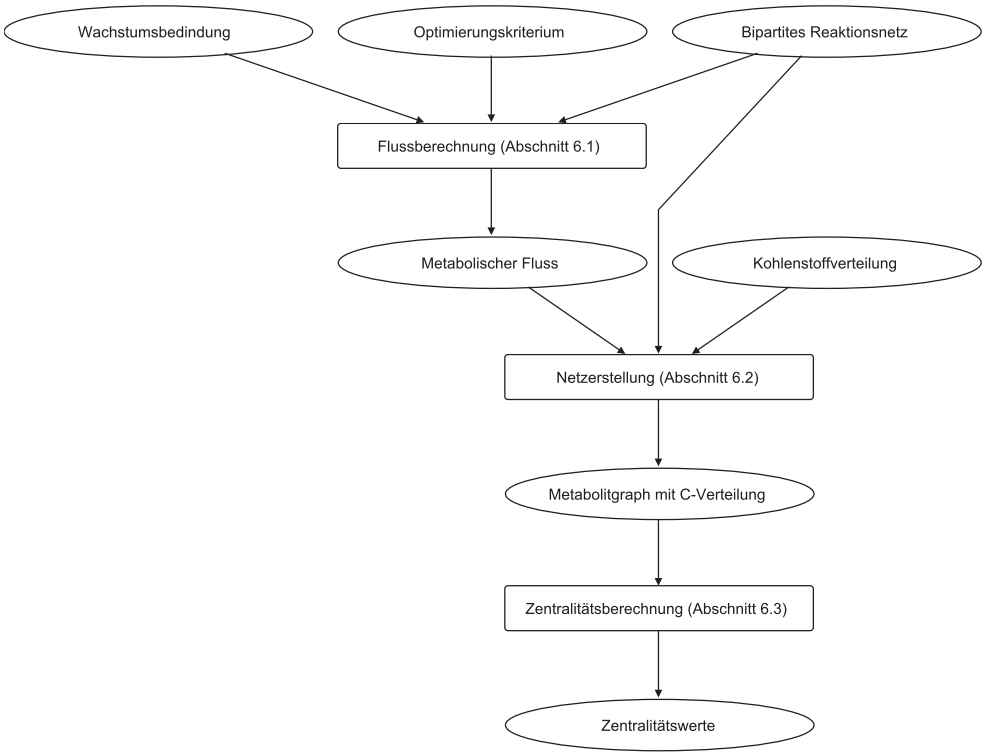


Abbildung 6.1.: Das Vorgehen zur Berechnung von Zentralitätswerten für Metaboliten auf der Basis der Fluss-basierten Zentralität. Ovale Knoten bezeichnen Daten und rechteckige Knoten die notwendigen Prozesse. Die dargestellten Prozesse werden in den benannten Abschnitten beschrieben.

Metaboliten und der innerhalb der Zelle vorherrschenden Bedingungen. Die Gesamtheit aller Bedingungen, die einen Einfluss auf die Höhe des Flusses haben, wird *Wachstumsbedingung* genannt.

Neben der Möglichkeit den Fluss über eine Reaktion experimentell zu ermitteln, besteht auch die Möglichkeit, diesen zu berechnen. Hierzu wird das Verhalten der Reaktion durch eine Differentialgleichung, der *Reaktionskinetik*, beschrieben. Damit diese Gleichung aufgestellt und gelöst werden kann, sind eine Reihe von Parametern, die *Reaktionsparameter*, erforderlich. Diese Parameter sind für jeden Organismus und jede Reaktion unterschiedlich und sie variieren auch unter verschiedenen Bedingungen innerhalb desselben Organismus. Momentan sind allerdings nur für eine geringe Anzahl von Reaktionen die Reaktionsmechanismen und -parameter überhaupt bekannt [66, 145]. Die Flüsse für alle Reaktionen eines Organismus anhand der Reaktionskinetiken zu berechnen ist deshalb momentan nicht bzw. nur sehr eingeschränkt möglich.

Eine Alternative zur Ermittlung der Flüsse über Messungen oder über die Berechnung auf der Basis von Differentialgleichungen besteht in der Anwendung der *Flux Balance Analysis* (*FBA*). Bei dieser Analyse werden ausschließlich Informationen aus der Stöchiometrie der Reaktionen verwendet. Dennoch kann durch diese Analyse eine hinreichend genau Bestimmung der Flüsse erreicht werden. Da die *Flux Balance Analysis* eine wichtige Möglichkeit zur Bestimmung von Flüssen für große Reaktionsnetze ist und da diese Methode auch für das im Abschnitt 6.5 analysierte Netzwerk eingesetzt wurde, wird diese im Folgenden kurz

zusammengefasst. Im Detail ist die FBA im Buch von Palsson beschrieben [133].

Eine biochemische Reaktion, beispielsweise die durch das Enzym Hexokinase katalysierte Umwandlung von α-D-Glukose in α-D-Glukose-6-phosphat (siehe Abbildung 6.2 (links)), kann als Gleichung (1 α-D-Glukose + 1 ATP $\Rightarrow$ 1 α-D-Glukose-6-phosphat + 1 ADP) dargestellt werden. Bei dieser Darstellung wird die Anzahl der durch die Reaktion umgesetzten Moleküle, die *Stöchiometrie*, vor die einzelnen Metabolitenbezeichnungen geschrieben. Um eine Reaktion eindeutig zu beschreiben, ist somit das Wissen über die Anzahl der umgesetzten Metaboliten und die Unterscheidung der Rolle (Substrat bzw. Produkt) des Metaboliten innerhalb der Reaktion notwendig. Die Anzahl der umgesetzten Metaboliten entspricht dabei genau dem entsprechenden stöchiometrischen Koeffizienten und für die Rolle gilt die Konvention, dass eines auf der linken (bzw. rechten) Seite der Reaktion genanntes Metabolit mit einem negativen (bzw. positiven) Vorzeichen notiert wird. Die obige Reaktionsgleichung lässt sich folglich durch die Notierung einer -1 für die beiden auf der linke Seite stehenden Metaboliten (α-D-Glukose und ATP) und einer 1 für die Metaboliten α-D-Glukose-6-phosphat und ADP eindeutig beschreiben.

Innerhalb einer Zelle bedingt ein metabolischer Fluss eine Änderung der Konzentrationen der an der Reaktion beteiligten Metaboliten. Die *Konzentrationsänderungen* der an einer Reaktion beteiligten Metaboliten x_i über die Zeit (geschrieben als $\dot{x_i} := \frac{dx_i}{dt}$) wird durch die Multiplikation des stöchiometrischen Koeffizienten s_i des i-ten an der Reaktion beteiligten Metaboliten mit dem Fluss v der Reaktion beschrieben ($\dot{x_i} = s_i v$). Mit der obigen Konvention werden folglich die Konzentrationsänderungen an den Metaboliten ATP und ADP, die durch die durch das Enzym Hexokinase katalysierte Reaktion vollzogen wird als $\dot{x_{\text{ATP}}} = -1 v_{\text{HK}}$ bzw. $\dot{x_{\text{ADP}}} = 1 v_{\text{HK}}$ notiert. Die tatsächliche Rolle eines Metaboliten, d.h. die Eigenschaft Substrat bzw. Produkt der Reaktion zu sein, ist dabei allerdings abhängig von der „Richtung“, in der eine Reaktion abläuft. Wenn der Fluss über eine Reaktion ein positives Vorzeichen hat, dann werden die Metaboliten auf der „linken Seite“ als Substrate und die Metaboliten auf der „rechten Seite“ als Produkte bezeichnet; für einen Fluss mit negativem Vorzeichen gilt das Umgekehrte.

Ein Reaktionsnetzwerk mit m Metaboliten und n Reaktionen kann, wie schon eine einzelne Reaktion, durch seine Stöchiometrie (genauer die *stöchiometrische Matrix* $S \in \mathbb{R}^{m \times n}$) eindeutig beschrieben werden. Die Komponenten s_{ij} dieser Matrix sind dabei die stöchiometrischen Koeffizienten des i-ten Metaboliten bzgl. der j-ten Reaktion. Entsprechend der Formel zur Berechnung der Konzentrationsänderung bedingt durch eine Reaktion werden die Konzentrationsänderungen aller Metaboliten innerhalb des Netzes durch die Multiplikation der Matrix mit dem Vektor der Flüsse aller Reaktionen beschrieben ($\dot{\vec{x}} = S\vec{v}$). Falls keine Konzentrationsänderungen auftreten, d.h. falls $\dot{\vec{x}} = \vec{0}$ gilt, dann befindet sich das betrachtete System im *Gleichgewichtszustand* (engl. *Steady state*).

Auch wenn sich das betrachtete System im Gleichgewichtszustand befindet, können die einzelnen Reaktionen einen Fluss aufweisen. Dieser Fall wird durch das Gleichungssystem $S\vec{v} = \vec{0}$ beschrieben. Da im Allgemeinen $m > n$ gilt, d.h. das im betrachteten System mehr Reaktionen als Metaboliten vorhanden sind, ist dieses Gleichungssystem i.d.R. nicht eindeutig lösbar. Durch das Festlegen einer zu optimierenden Bedingung und Anwendung der lineare oder quadratischen Optimierung, lässt sich in der Regel dennoch ein Flussvektor $\vec{v}$ ermitteln. Als Optimierungskriterium wird dabei üblicherweise verlangt, dass der modellierte Organismus alle zur Verfügung stehenden Ressourcen zum Wachstum nutzen soll (die so genannte Maximierung der Biomasse) oder das vom Organismus möglichst wenig Energie (in Form von ATP) verbraucht werden soll [154]. Um spezifische Eigenschaften biochemischer Reaktionsnetze in der Optimierung zu berücksichtigen, werden üblicherweise zusätzliche

Bedingungen festgelegt. Hierzu gehört beispielsweise, dass einige Reaktionen nur in eine Richtung ablaufen können, und folglich der entsprechende Flusswert nur positiv bzw. nur negativ sein kann, oder das nur eine gewisse Menge an Sauerstoff für die Atmung der Zellen zur Verfügung steht. Durch die Änderung dieser Bedingungen können folglich unterschiedliche Wachstums- bzw. Umgebungsbedingungen für einen modellierten Organismus simuliert werden.

Mit Hilfe der FBA lässt sich somit für ein gegebenes Reaktionsnetz der metabolische Fluss für alle Reaktionen ermitteln. Dabei wird das festgelegte Optimierungskriterium berücksichtigt und unterschiedliche Wachstumsbedingungen können simuliert werden. In der Abbildung 6.1 ist dieser Teilschritt bei der Berechnung der Zentralitätswerte für die Metaboliten eines Reaktionsnetzes entsprechend dargestellt.

Der mittels FBA (oder einer vergleichbaren Methode) ermittelte Flussvektor $\vec{v}$ kann als Knotengewicht für die Knoten der Menge der Reaktionen des bipartiten Reaktionsnetzes interpretiert werden. Durch die Hinzunahme der Flussinformationen wird folglich aus dem bipartiten Graphen ein Graph, in dem die Reaktionsknoten mit Flüssen und die Metaboliten mit dem konstanten Gewicht 1 markiert sind. Zusätzlich zu den quantitativen Flusswerten werden durch die FBA die Richtungen, in der die Reaktionen ablaufen, ermittelt. Bereits durch die Verwendung dieser Information unterscheidet sich die Analyse mit Hilfe der im Folgenden beschriebenen Zentralität von den bisherigen Analysen: Vorherige Analysen von Reaktionsnetzen haben entweder ungerichtete Netze verwendet oder angenommen, dass viele der Reaktionen gerichtet sind und dabei in beide Richtungen ablaufen können (siehe Abschnitt 4.2). Durch die Hinzunahme der Flussinformationen ist es somit möglich, das Netzwerk als gerichtetes Netzwerk zu betrachten. Da die Information über die Richtung der Reaktion auf dem Ergebnis der Flussermittlung basieren, sind diese in sich konsistent.

Im Folgenden wird aus dem bipartiten Graph mit Knotengewichten ein, den Kohlenstofffluss innerhalb des modellierten Systems beschreibender, unipartiter Graph mit Kantengewichten abgeleitet.

6.2. Der Metabolitgraph mit Kohlenstofffluss

Kohlenstoff ist, neben Wasser, der zweite Grundbaustein des Lebens. Für das Verständnis von Stoffwechselvorgängen kann deshalb die Betrachtung der innerhalb eines Organismus durch biochemische Reaktionen umgesetzten Menge an Metaboliten auf die Betrachtung der Menge des in den Metaboliten enthaltenen Kohlenstoffs reduziert werden. Die durch eine biochemische Reaktion umgesetzte Anzahl an Kohlenstoffatomen wird im Folgenden als *Kohlenstofffluss der Reaktion* bzw. als *Kohlenstoffverteilung der Reaktion* bezeichnet.

Aus jedem bipartiten Reaktionsgraphen lassen sich zwei unipartite Graphen, der Metabolit- und der Reaktionsgraph, ableiten (siehe Abschnitt 3.4.2). Im Folgenden wird, da eine Zentralität für Metaboliten beschrieben werden soll, die Erstellung eines den Kohlenstofffluss innerhalb des Organismus widerspiegelnden Metabolitgraphen zu einem gegebenen bipartiten Graphen mit Flussinformationen beschrieben. Der in diesem Abschnitt beschriebene Aufbereitungsprozess ist in der Abbildung 6.1 in der Bildmitte dargestellt und wird im Abschnitt 6.4.1 (Algorithmus 6.1 und 6.2) formalisiert.

Üblicherweise wird eine als bipartiter Graph modellierte Reaktion durch die in der Bildmitte in Abbildung 6.2 dargestellte Transformation in einen unipartiten Metabolitgraphen überführt. Bei dieser Transformation wird jedes Substratmetabolit mit jedem Produktmetabolit verbunden. Problematisch an dieser Art der Umwandlung ist, dass dabei auch Kanten

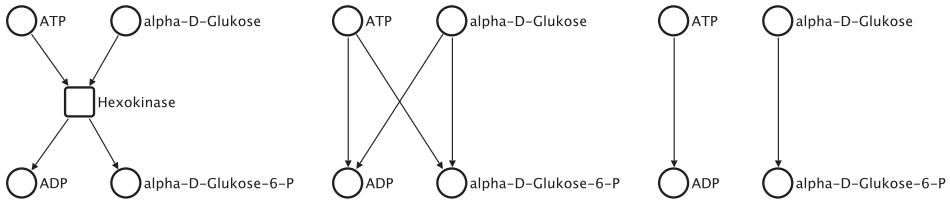


Abbildung 6.2.: (links) Die durch das Enzym Hexokinase katalysierte Reaktion, dargestellt als bipartiter Graph. (mitte) Der nach der üblicherweise verwendeten Transformation daraus abgeleitete Metabolitgraph. Problematisch an dieser Transformation ist, dass je eine Kante zwischen den Metaboliten α-D-Glukose und ADP bzw. ATP und α-D-Glukose-6-phosphat erstellt wird, obwohl in dieser Reaktion zwischen diesen Metaboliten keine Kohlenstoffatome ausgetauscht werden. (rechts) Der Metabolitgraph, der entsteht, wenn nur der Kohlenstoffffluss bei der Verbindung der Metaboliten berücksichtigt wird.

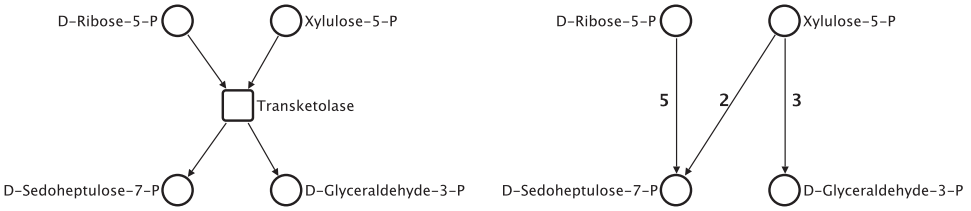


Abbildung 6.3.: Die durch das Enzym Transketolase katalysierte Reaktion als bipartiter Graph (links) und als Metabolitgraph unter Berücksichtigung des tatsächlichen Kohlenstoffflusses (rechts). Die Kantengewichte geben die Anzahl der in der Reaktion übertragenen Kohlenstoffatome an.

zwischen Metaboliten erstellt werden, die in der betrachteten Reaktion keine Kohlenstoffatome miteinander austauschen.

Die in der Abbildung 6.2 auf der rechten Seite dargestellte Umwandlung derselben Reaktion berücksichtigt bei der Erstellung der Kanten den Kohlenstoffffluss zwischen den Metaboliten. Zwei Metaboliten werden nur dann miteinander verbunden, wenn mindestens ein Kohlenstoffatom vom Substrat- zum Produktmetaboliten überführt wird. In der gezeigten Reaktion findet eine Übertragung von Kohlenstoffatomen nur zwischen den Metaboliten α-D-Glukose und α-D-Glukose-6-phosphat bzw. ATP und ADP statt. Folglich sind bei dieser Umwandlung des bipartiten Graphen in den unipartiten Graphen nur die entsprechenden Metaboliten miteinander verbunden.

Einige biochemische Reaktionen, beispielsweise die durch die Enzyme Transketolase (siehe Abbildung 6.3(links)) oder Transaldolase katalysierten, zerlegen eines der Substratmetaboliten und verbinden ein abgespaltenes Teilmolekül des Substratmetaboliten mit dem anderen Substratmetabolit. Ein Austausch von Kohlenstoffatomen findet bei dieser Art von Reaktion folglich nur zwischen einigen Kombinationen von Metaboliten statt. Die Anzahl der übertragenen Atome kann dabei für jedes Paar von Metaboliten unterschiedlich sein. Im Folgenden wird deshalb die Anzahl der übertragenen Kohlenstoffatome innerhalb des erstellten Metabolitgraphen als Kantengewicht an die betreffende Kante notiert. Für die durch das Enzym Transketolase katalysierte Reaktion ist der zugehörige Metabolitgraph mit Kantengewichten in der Abbildung 6.3 (rechts) dargestellt.

Als Grundlage für die Berechnung der weiter unten beschriebenen Zentralität werden alle

Reaktionen eines zu untersuchenden bipartiten Reaktionsnetzes mit Flussvektor entsprechend der vorigen Beschreibung transformiert (formalisiert im Algorithmus 6.1, Seite 80). Zusätzlich wird die Gesamtmenge der durch eine Reaktion zwischen den Substrat- und Produktmetaboliten übertragenen Kohlenstoffatomen durch Multiplikation des Reaktionsflusses mit der Anzahl der übertragenen Kohlenstoffatome je Kante ermittelt und als Kantengewicht an die entsprechenden Kanten des Metabolitgraphen notiert. Der durch diese Transformation entstehende kantengewichtete Metabolitgraph wird im Folgenden *Metabolitgraph mit Kohlenstoffffluss* (*MGKF*) genannt.

Bei der Erstellung des MGKF wird die ermittelte Fließrichtung aller Reaktionen berücksichtigt. Folglich wird, wenn eine Übertragung von Kohlenstoffatomen erfolgt, zwischen zwei Metaboliten eine Kante entsprechend der ermittelten Fließrichtung eingefügt. Eine bereits existierende Kante wird dabei dahingehend berücksichtigt, dass das berechnete Kantengewicht zum Kantengewicht der bereits existierenden Kante hinzu addiert wird.

Antiparallele Kanten, die nach der Aufbereitung aller Reaktionen eventuell entstandenen sind, beschreiben einen wechselseitigen Kohlenstoffffluss zwischen den beteiligten Metaboliten. Dieser wechselseitige Fluss findet in der Zelle allerdings nicht statt. Folglich wird innerhalb des Modells nur der Nettofluss zwischen den beteiligten Metaboliten berücksichtigt und die antiparallelen Kanten werden entsprechend reduziert. Hierzu wird die Kante mit dem geringeren Kantengewicht gelöscht und das Kantengewicht der dazugehörigen antiparallelen Kante um das Kantengewicht der gelöschten Kante verringert. Für den Fall, dass beide antiparallelen Kanten exakt denselben Kohlenstoffffluss aufweisen, werden, da kein Nettofluss zwischen den Metaboliten im Organismus stattfindet, beide Kanten gelöscht (siehe auch Algorithmus 6.2).

Der so entstandene MGKF ist ein gerichteter Graph mit Kantengewichten. Dieser Graph beschreibt den Kohlenstoffffluss innerhalb des modellieren Systems korrekt. Einzelne Metaboliten ohne Nettofluss können in diesem Graph von der größten Zusammenhangskomponente getrennt sein. Aufgrund der Netzwerkstruktur metabolischer Reaktionsnetze enthält der Graph allerdings üblicherweise eine sehr große schwache Zusammenhangskomponente [110]. Für die Berechnung der Zentralitätswerte der Metaboliten wird im Folgenden nur noch diese größte schwache Zusammenhangskomponente betrachtet und weiterhin als MGKF bezeichnet.

6.3. Ein Zentralitätsmaß für Metabolitgraphen mit Kohlenstoffffluss

Jedes Zentralitätsmaß benötigt ein eindeutiges Bewertungskriterium anhand dem entschieden wird, wann ein Knoten als „zentraler“ zu bezeichnen ist als ein anderer Knoten. Für die *In-Degree*-Zentralität gilt beispielsweise, dass ein Knoten v_1 „zentraler“ als ein Knoten v_2 ist, wenn die Nachbarschaft von v_1 (im Sinne der Kardinalität) größer ist als die Nachbarschaft von v_2 ($|\Gamma^-(v_1)| > |\Gamma^-(v_2)|$).

Ein Zentralitätsmaß für die Metaboliten eines Reaktionsnetzes benötigt folglich ebenfalls ein solches Bewertungskriterium. Das hier vorgestellte Zentralitätsmaß verwendet als Bewertungskriterium das *Potential eines Metaboliten*. Das Potential eines Metaboliten bezeichnet dabei die Gesamtmenge der innerhalb eines Organismus aus einem betrachteten Metaboliten m_1 herstellbaren Menge anderer Metaboliten. Ein Metabolit m_1 gilt somit als „zentraler“ als ein Metabolit m_2, wenn aus dem Metaboliten m_1 eine größere Gesamtmenge an Metaboliten hergestellt werden kann als aus dem Metaboliten m_2.

Auf der Basis eines Metabolitgraphen mit Kohlenstoffffluss lässt sich die aus einem Metaboliten ableitbare Gesamtmenge an erzeugbaren Metaboliten ermitteln und somit eine Zentralität definieren, die dem obigen Kriterium entspricht. Hierzu werden die Kantengewichte des MGKF als Kapazitäten für ein *Max-Flow*-Problem interpretiert und es wird für jedes Metabolit (über die Lösung eines *Max-Flow*-Problems) die Menge der aus diesem Metabolit erzeugbaren Metaboliten ermittelt.

Definiert wird diese Zentralität wie folgt:

Definition 6.1 (*Max-Flow-Closeness*-Zentralität)
Sei (G, c) ein Metabolitgraph mit Kohlenstoffffluss, wobei c das durch Algorithmus 6.2 ermittelte Kantengewicht bezeichnet. Die Max-Flow-Closeness-Zentralität *[94] ist dann definiert als*

$$\mathcal{C}_{mfc}(s) := \sum_{t \in V(G) \setminus \{s\}} \text{MaxFlowValue}(G, c, s, t)$$

Hierbei bezeichnet $\text{MaxFlowValue}(G, c, s, t)$ *den Wert des maximalen Flusses von s nach t im gewichteten Graphen (G, c) (siehe Definition 2.18).*

Die Bezeichnung der Zentralität als *Closeness* erfolgt dabei in Anlehnung an die bereits existierenden *Closeness*-Zentralitäten *Current-Flow Closeness* und *Shortest-Path Closeness*. Beide existierenden Zentralitäten summieren einen nur vom betrachteten Knoten s, von allen anderen Knoten $t \neq s$ und dem betrachteten Graph G abhängigen Wert auf; die *Current-Flow Closeness* beispielsweise summiert über die effektiven Widerstände aller Knotenpaare $(s, t) \in (V \times V)$. Etwas präziser könnte die Zentralität als *Out-Max-Flow-Closeness*-Zentralität bezeichnet werden, da auch eine Definition über den eingehenden Fluss (*In-Max-Flow-Closeness*-Zentralität) denkbar ist.

Mit Hilfe der *Max-Flow-Closeness*-Zentralität können die Knoten eines metabolischen Reaktionsnetzes mit Flussinformationen entsprechend ihres Potentials in eine Reihenfolge gebracht werden.

Im folgenden Abschnitt werden die Algorithmen zur Berechnung der vorgestellten Zentralität eingeführt und im darauf folgenden Abschnitt wird die Zentralität zur Analyse eines metabolischen Reaktionsnetzes von *Escherichia coli* eingesetzt.

6.4. Algorithmen für die Metabolitgrapherstellung und die *Max-Flow-Closeness*-Zentralität

Die für die Erstellung des MGKF und die Berechnung der Zentralitätswerte notwendigen Algorithmen werden in diesem Abschnitt beschrieben.

6.4.1. Erstellung des Metabolitgraph mit Kohlenstoffffluss

Die Erstellung des MGKF erfolgt in einem zweistufigen Verfahren. Zuerst (Algorithmus 6.1) wird aus dem bipartiten Graphen, der Kohlenstoffverteilung der Reaktionen und dem Flussvektor ein unipartiter Graph erstellt. Im zweiten Schritt (Algorithmus 6.2) werden aus diesem Graph eventuelle existierende antiparallele Kanten entfernt.

In den Tabellen 6.2 und 6.1 sind die in den beiden Algorithmen verwendeten Datenstrukturen, Funktionen und die zugehörigen Komplexitätsabschätzungen aufgeführt. Für die Betrachtung der Laufzeitkomplexität wird im Folgenden vorausgesetzt, dass der Flussvektor als Liste, alle Mengen als AVL-Bäume und Graphen mittels Adjazenzlisten implementiert

Algorithmus 6.1 Erstellung eines Metabolitgraph mit Kohlenstoffffluss (MGKF) aus einem gegebenen bipartiten Reaktionsnetz, einem dazugehörigen Flussvektor und einer Verteilung der Kohlenstoffatome innerhalb der betrachteten Reaktionen.

Eingabe: $G1$: Bipartites Reaktionsnetz, $\vec{v}$: Flussinformation bestehend aus Reaktionsname und Flusswert, $C1$: Tabelle mit der Verteilung der Kohlenstoffatome zu den in $G1$ modellierten Reaktionen

Ausgabe: $G2$: Metabolitgraph mit Kohlenstoffffluss (MGKF)

```
1: G2 ← createEmptyGraph
2: for all fluxInfo ∈ v⃗ do
3:     flux ← fluxInfo.fluxValue
4:     if flux = 0 then
5:         Continue
6:     reactionName ← fluxInfo.reactionName
7:     if flux > 0 then
8:         substrateNames ← getSubstrateNames(G1, reactionName)
9:         productNames ← getProductNames(G1, reactionName)
10:    else
11:        substrateNames ← getProductNames(G1, reactionName)
12:        productNames ← getSubstrateNames(G1, reactionName)
13:        flux ← −flux
14:    for all substrateName ∈ substrateNames do
15:        substrateVertex ← getOrCreateVertex(G2, substrateName)
16:        for all productName ∈ productNames do
17:            productVertex ← getOrCreateVertex(G2, productName)
18:            carbonTransfer ←
                   getCarbonTransfer(C1, reactionName, substrateName, productName)
19:            if carbonTransfer = 0 then
20:                Continue
21:            carbonFlux ← flux ∗ carbonTransfer
22:            edge ← findEdge(G2, substrateVertex, productVertex)
23:            if edge = Null then
24:                createEdge(G2, substrateVertex, productVertex, carbonFlux)
25:            else
26:                setEdgeWeight(edge, getEdgeWeight(edge)+carbonFlux)
```

sind [155, 156]. Der Datentyp Kohlenstoffverteilung ist eine Tabelle, in der die Anzahl der innerhalb einer Reaktion umgesetzten Kohlenstoffatome angegeben sind. Für die durch das Enzym Transketolase katalysierte Reaktion (siehe Abbildung 6.3) ist diese Verteilung der Kohlenstoffatome exemplarisch in der Tabelle 6.3 angegeben. Implementiert werden kann diese Tabelle als mehrfach verschachtelte Hash-Struktur oder als mehrfach verschachtelter balancierter Binärbaum. In der Programmiersprache Java könnte sie beispielsweise als `TreeMap<String, TreeMap<String, TreeMap<String, Double>>>` deklariert werden, wobei die einzelnen Zeichenketten den Reaktionsbezeichner und die beiden Metabolitnamen enthalten und die Anzahl der übertragenen Kohlenstoffatome im `Double` abgelegt wird. Die Größe der Datenstruktur ist nach oben durch $|R| * |M|^2$, d.h. durch die Anzahl der Reaktionen und das Quadrat der Anzahl der Metaboliten innerhalb des zu untersuchenden Reaktionsnetzes, begrenzt.

Für den Algorithmus 6.1 ergibt eine sehr grobe Komplexitätsabschätzung eine obere Schranke für die Laufzeit von $\mathcal{O}(|V|^4 + |V|^3 * |E|)$ und für den Algorithmus 6.2 liegt diese obere Schranke bei $\mathcal{O}(|V| * |E| + |E|^2)$. Diese Abschätzungen sind allerdings deutlich zu

Algorithmus 6.2 Entfernung von antiparallelen Kanten aus einem Metabolitgraph mit Kohlenstoffffluss

Eingabe: G: Metabolitgraph mit Kohlenstoffffluss (MGKF)
Ausgabe: G: MGKF ohne antiparallele Kanten

```
1: untreatedEdges ← getAllEdges(G)
2: while untreatedEdges ≠ ∅ do
3:     edge ← getOneElement(untreatedEdges)
4:     apEdge ← getAntiParallelEdge(G, edge)
5:     if apEdge = Null then
6:         removeElement(untreatedEdges, edge)
7:     else
8:         weight ← getEdgeWeight(edge)
9:         apWeight ← getEdgeWeight(apEdge)
10:        if weight = apWeight then
11:            deleteEdge(G, edge)
12:            deleteEdge(G, apEdge)
13:        else if (weight > apWeight) then
14:            setEdgeWeight(edge, weight − apWeight)
15:            deleteEdge(G, apEdge)
16:        else
17:            setEdgeWeight(apEdge, apWeight − weight)
18:            deleteEdge(G, edge)
19:        removeElement(untreatedEdges, edge)
20:        removeElement(untreatedEdges, apEdge)
```

Algorithmus 6.3 Berechnung der Zentralitätswerte auf der Basis eines Metabolitgraphen mit Kohlenstoffffluss (MGKF).

Eingabe: G: Metabolitgraph mit Kohlenstoffffluss, c: Kantengewichte des MGKF
Ausgabe: Zentralitätswerte $\mathcal{C}_{mfc}(v)$ für alle Knoten $v \in V(G)$

```
1: for all s ∈ V(G) do
2:     C_mfc(s) ← 0
3:     for all t ∈ V(G) \ {s} do
4:         C_mfc(s) ← C_mfc(s) + MaxFlowValue(G, c, s, t)
```

grob, da eine wesentliche Eigenschaft der biochemischen Reaktionsnetze, nämlich das es sich um bipartite Graphen mit nur wenigen Vorgängern bzw. Nachfolgern pro Reaktion handelt, nicht ausgenutzt wird. Für das im Abschnitt 6.5 zu untersuchende Reaktionsnetz von *E. coli* gilt beispielsweise, dass der durchschnittliche kombinierte Eingangs- und Ausgangsgrad bei $5,4$ liegt. Die drei größten Werte sind 36 (1x), 30 (1x) und 20 (9x). Folglich muss zwar in der $\mathcal{O}$-Notation in den Zeilen 14 und 16 des Algorithmus 6.1 von einer Komplexität von $\mathcal{O}(|V|^2)$ ausgegangen werden, dieser Wert wird aber in der Regel jeweils deutlich unterschritten.

6.4.2. Berechnung der Zentralitätswerte auf der Basis des MGKF

Zur Berechnung der Zentralitätswerte der Knoten eines MGKF ist jeweils der maximale s-t-Fluss für alle Knotenpaare (s,t) zu bestimmen und zu summieren, siehe Algorithmus 6.3. Die Berechnung eines maximalen s-t-Flusses kann beispielsweise mit dem Algorithmus von Goldberg und Tarjan in einer Laufzeit von $\mathcal{O}(|V||E|\log(|V|^2/|E|))$ erfolgen [1, 64]. Hieraus folgt unmittelbar die obere Schranke für die Anzahl der Berechnungsschritte der Zentralitätswerte für die Knoten des MGKF von $\mathcal{O}(|V|^2 * |V||E|\log(|V|^2/|E|))$.

Eine Alternative zur Verwendung eines Algorithmus für die Berechnung der einzelnen *Max-Flow*-Werte besteht in der Verwendung eines spezifischen Algorithmus für die Lösung eines *All-Pairs* Problems. Für dünnbesetzte Graphen wurde ein *All-Pairs Min-Cut* Algorithmus von Arikati, Chaudhuri und Zaroliagis vorgeschlagen. Dieser Algorithmus berechnet jeweils den minimalen Schnitt[1] zwischen allen Knotenpaaren in einer Laufzeit von $\mathcal{O}(|V|^2 + \gamma^4 \log \gamma)$. Mit γ wird dabei die Anzahl der *Hammocks*, der kantendisjunkten kreisartig planaren Teilgraphen, von G bezeichnet. Dieser Wert kann mit $1 \leq \gamma \leq \Theta(|V|)$ abgeschätzt werden [13]. Hierbei bezeichnet $\Theta(n)$ die scharfe Abschätzung der asymptotischen Komplexität, im Unterschied zur Notation $\mathcal{O}$, bei der nur eine obere Schranke angegeben wird.

6.5. Analyse eines metabolischen Reaktionsnetzes von *Escherichia coli* mittels der Fluss-basierten Zentralität

In diesem Abschnitt wird das in den vorigen Abschnitten vorgestellte Zentralitätsmaß für Metaboliten exemplarisch auf ein Netzwerk angewendet. Hierbei wird, wie bereits für die Motiv-basierte Zentralität (siehe Abschnitt 5.4), ein Netzwerk für den Organismus *Escherichia coli* verwendet. Im Unterschied zum Abschnitt 5.4 wird allerdings ein biochemisches Reaktionsnetz (siehe Abschnitt 3.4.2) analysiert.

Zur Berechnung der Zentralitätswerte für die Metaboliten wird, gemäß Abbildung 6.1, ein bipartites Reaktionsnetz, eine Wachstumsbedingung und ein Optimierungskriterium benötigt. Aus diesen drei Informationen und der Kohlenstoffverteilung für die Reaktionen des zu analysierenden Netzes kann dann ein Metabolitgraph mit Kohlenstoffffluss (MGKF) ermittelt und ein Zentralitätsvektor berechnet werden. Um den Einfluss unterschiedlicher Wachstumsbedingungen auf die entstehende Reihenfolge der Metaboliten zu veranschaulichen, werden im Folgenden mehrere Wachstumsbedingungen betrachtet. Für jede dieser Wachstumsbedingungen wird dazu ein Zentralitätsvektor bestimmt und somit eine Reihenfolge der Metaboliten ermittelt.

Im Detail gliedert sich die Analyse des gegebenen Netzes in vier Teilschritte: Als erstes werden für das gegebene Reaktionsnetz insgesamt 19 verschiedene Wachstumsbedingungen beschrieben und die zugehörigen Flussinformationen mittels FBA bestimmt. Danach wird auf der Basis dieser Flussvektoren jeweils ein zugehöriger MGKF erstellt und anhand diesem der jeweilige Zentralitätsvektor ermittelt. Diese Zentralitätsvektoren werden dann analysiert und mit bereits bestehenden Resultaten basierend auf anderen Zentralitätsmaßen verglichen. Als letztes werden die Zentralitätswerte zur Modularisierung des Reaktionsnetzes in Stoffwechselwege eingesetzt.

6.5.1. Metabolischer Fluss für 19 Wachstumsbedingungen

Im Folgenden wird das von J. Reed und Koautoren vorgestellte Modell des Stoffwechsels von *E. coli K-12* [142] für die Demonstration der Fluss-basierten Zentralität verwendet. Dieses Modell umfasst insgesamt 904 Metaboliten und 931 Reaktionen. In der Abbildung 6.4 ist der Metabolitgraph des Zentralstoffwechsels für *E. coli*, bestehend aus dem Zitronensäurezyklus, der Glykolyse und dem Pentosephosphat-Stoffwechselweg, basierend auf diesem Modell, dargestellt.

[1] Der Wert eines maximalen s-t-Flusses ist gleich die Kapazität eines kleinsten s-t-Schnittes im betrachteten Graphen [55].

Auf der Basis dieses Modells wurden insgesamt 19 Wachstumsbedingungen simuliert. Die wesentlichen Eigenschaften dieser Wachstumsbedingungen sind in der Tabelle 6.4 zusammengefasst. Die erste Spalte dieser Tabelle enthält die im Folgenden verwendete Bezeichnung der Wachstumsbedingung, die zweite Spalte den Namen und die Summenformel des Metaboliten, der dem (virtuellen) Organismus zugeführt wurde, die dritte Spalte bezeichnet die dem Organismus bereitgestellte Menge des vorgenannten Metaboliten und die vierte Spalte gibt an, ob dem Organismus Sauerstoff zur Verfügung stand (sogenanntes aerobes bzw. anaerobes Wachstum). Um die Ergebnisse der Zentralitätsanalyse vergleichbar zu machen, wurde die Menge des verfügbaren Substrats, z.B. Azetat oder Glukose, für die einzelnen Wachstumsbedingungen auf das Äquivalent von $60 \frac{\text{mmol Kohlenstoff}}{\text{h}} \frac{\text{g}}{\text{DW}}$ normiert.

Die weiteren Einstellungen für die Berechnungen des metabolischen Flusses für die Wachstumsbedingungen erfolgten wie bei der Vorstellung des Modells iJR904 vorgeschlagen. Das heißt, die Metaboliten Kohlenstoffdioxid, Ammoniak, Schwefel, Natrium, Kalium, Phosphor, Protonen, Wasser und Eisen (II) können in beliebiger Menge aufgenommen und abgegeben werden. Die anderen im Modell als „extern“ markierten Metaboliten dürfen hingegen nur an die Umgebung abgegeben werden. Im Fall, dass das Wachstum unter Entzug von Sauerstoff simuliert wurde (anaerobes Wachstum), wurde die Menge des aufgenommenen Sauerstoffs auf 0 fixiert. Für aerobes Wachstum war die Menge des verfügbaren Sauerstoffs nicht limitiert. Zusätzlich wurde die im Modell vorgesehene Reaktion *ATP maintenance* auf $7,6 \frac{\text{mmol}}{\text{h}} \frac{\text{g}}{\text{DW}}$ fixiert [142].

Das gewählte Optimierungskriterium für die Optimierung im Rahmen der durchgeführten FBA ist die so genannte Maximierung der Biomasse. Hierbei wird angenommen, dass der simulierte Organismus alle verfügbaren Möglichkeiten ausnutzt, um zu „wachsen“. Dieses „Wachstum“ wird dabei dadurch simuliert, dass eine Reaktion mit dem Namen „Biomasse“ in das Modell aufgenommen wird. Diese Reaktion erhält als Eingangssubstrate jeweils die Bestandteile der Biomasse und als Produkt ein Metabolit mit dem Namen „Biomasse“. Bei der Optimierung im Rahmen der FBA wird dann der Fluss über diese Reaktion maximiert.

Berechnet wurde die FBA mit Hilfe der COBRA Toolbox [19]. Hierzu wurde für alle 19 Wachstumsbedingungen zuerst der Fluss über die Reaktion Biomasse maximiert. Dabei kam der COIN-OR Linear Program Solver (CLP solver) [108] zum Einsatz. Im Anschluss daran erfolgte ein zweiter Optimierungsschritt mittels quadratischer Optimierung, wiederum mit Hilfe des CLP Solvers. Durch diesen zweiten Optimierungsschritt konnten Probleme mit mehrfach auftretenden alternativen Lösungen vermieden werden, denn während des zweiten Optimierungsschritts wurde der Biomassefluss auf den Wert des ersten Optimierungsschritts fixiert [115].

Als Resultat der FBA steht für jede Wachstumsbedingung schlussendlich ein Vektor mit Werten für den (simulierten) metabolischen Fluss innerhalb von *E. coli* unter den Einschränkungen der Wachstumsbedingung und der Forderung nach möglichst starkem Wachstum zur Verfügung.

6.5.2. Zentralitätsvektoren auf der Basis der Wachstumsbedingungen

Um zu den berechneten 19 Flussvektoren die entsprechenden Zentralitätsvektoren zu ermitteln, sind als Zwischenschritt, wie in Abbildung 6.1 dargestellt, jeweils die dazugehörigen Metabolitgraphen mit Kohlenstoffffluss (MGKFs) zu erstellen. Hierzu ist als Eingabe wiederum das bipartite Reaktionsnetz, die im vorigen Abschnitt beschriebenen Flussvektoren und zusätzlich die Kohlenstoffverteilung der einzelnen Reaktionen des Reaktionsnetzes erforderlich.

Der Aufbau der Tabelle mit der Kohlenstoffverteilung der einzelnen Reaktionen ist bereits exemplarisch in der Tabelle 6.3 dargestellt. Insgesamt besteht die Tabelle für das analysierte Modell von *E. coli* aus 2516 Einträgen für die 464 Reaktionen, die nach der Berechnung des metabolischen Flusses im vorigen Schritt in einer der betrachteten Wachstumsbedingungen einen Fluss ungleich 0 aufweisen. Erstellt wurde diese Tabelle unter Zuhilfenahme der Datenbank KEGG RPAIR [128]. In dieser Datenbank sind für die Substrat-Produkt-Paare von enzymkatalysierten Reaktionen jeweils die chemischen Strukturtransformationen beschrieben. Für die Reaktionen des betrachteten Reaktionsnetzes wurden folglich die entsprechenden Strukturtransformationen manuell aus der Datenbank extrahiert und in die gewünschte Tabellenstruktur überführt.

Basierend auf dem bipartiten Reaktionsnetz, den 19 Flussvektoren und der Kohlenstoffverteilung wurden die zur Berechnung der Zentralitätsvektoren erforderlichen 19 MGKFs erstellt. Diese Metabolitgraphen haben, bedingt durch die Tatsache, dass nicht alle Reaktionen einen Fluss aufweisen, eine Größe von 283–331 Metaboliten bei 488–555 Kanten. Wobei, wie schon in Abschnitt 6.2 erläutert, nur jeweils die größte schwache Zusammenhangskomponente betrachtet wurde. Auf der Basis dieser 19 MGKFs können dann die Zentralitätsvektoren für die Wachstumsbedingungen berechnet werden. Hierbei kommt der in Abschnitt 6.4.2 beschriebene Algorithmus zum Einsatz.

In der Tabelle 6.5 sind die Zentralitätswerte für die Metaboliten und die 19 Wachstumsbedingungen zusammengefasst. Dargestellt ist eine Liste der Top-30 Metaboliten, basierend auf der Summierung der Zentralitätswerte über die Wachstumsbedingungen hinweg. Es zeigt sich, dass die durch die Fluss-basierte Zentralitätsanalyse erlangte Reihenfolge der Metaboliten eine nahezu 100%ige Übereinstimmung mit den Metaboliten des Zentralstoffwechsels von *E. coli* erreicht. Im folgenden Abschnitt wird die Überlegenheit der Fluss-basierten Zentralität durch einen Vergleich mit Resultaten basierend auf anderen Zentralitätsmaßen noch deutlicher.

6.5.3. Vergleich der Resultate der Fluss-basierten Zentralität mit Resultaten basierend auf anderen Zentralitäten für metabolische Reaktionsnetze

Im Abschnitt 4.2.1 wurden die bereits existierenden Veröffentlichungen in denen Zentralitäten zur Bestimmung einer Reihenfolge von Metaboliten angewendet wurden vorgestellt und zusammengefasst. Im Folgenden werden die in diesen Publikationen erzielten Resultate mit den Resultaten der Fluss-basierten Zentralität verglichen. In der Tabelle A.1 auf Seite 108 sind hierzu die entsprechenden Resultate, jeweils als Top-10 Liste, aufgeführt.

Die von Jeong und Koautoren [79] ermittelte Reihenfolge von Metaboliten wird eindeutig von Kofaktoren dominiert. Bei der dort gewählten Kombination aus Netzwerkmodellierung und Zentralität (*Degree*) ist dieses aber nicht verwunderlich. Auch im hier verwendeten Modell iJR904 sind 19% der Reaktionen mit dem Metabolit ATP verbunden. Diese hohe Anzahl an Verbindungen zu Kofaktoren ist allerdings nicht überraschend und für alle Organismen und unabhängig von der vorherrschenden Wachstumsbedingung der Fall.

In den ermittelten Reihenfolgen von Wagner & Fell [50, 49, 167] dominieren Metaboliten des Zitronensäurezyklus, wiederum unabhängig von der betrachteten Wachstumsbedingung. Die in der Publikation von Jeong und Koautoren als wichtig identifizierten Metaboliten wurden bei Wagner & Fell bereits vor der Analyse aus dem betrachteten Netzwerk entfernt. Der Konstruktionsprozess bei Wagner & Fell ist allerdings vergleichbar zum Konstruktionsprozess von Jeong und Koautoren.

Ma & Zeng haben bei der Erstellung der zu analysierenden Reaktionsnetzwerke die Kofak-

toren (manuell) eliminiert und, soweit möglich, die Reaktionsrichtung der einzelnen Reaktionen berücksichtigt [110, 111]. Die angegebene Liste mit den Top-10 Metaboliten gemäß der *In-* und *Out-Closeness*-Zentralität für *E. coli* enthält acht Metaboliten aus den Stoffwechselwegen Glykolyse und Zitronensäurezyklus. Das Metabolit mit dem höchsten Zentralitätswert ist Pyruvat, das Metabolit am Übergang von der Glykolyse in den Zitronensäurezyklus.

Vergleichbar zur hier vorgestellten Idee des MGKF analysierte Arita einen Metabolitgraphen für *E. coli* [14]. Von Arita wurde hierzu die Information über die Verwendung der einzelnen Kohlenstoffatome innerhalb der Reaktionen ausgewertet und eine Verbindung zwischen zwei Metaboliten nur hergestellt, wenn mindestens ein Kohlenstoffatom ausgetauscht wurde. Im Unterschied zum MGKF wurde die Menge des übertragenen Kohlenstoffs allerdings nicht berücksichtigt. Konsequenterweise wird die Liste der Top-10 Metaboliten von Metaboliten angeführt, die Kohlenstoffatome an viele andere Metaboliten abgeben können, beispielsweise CO_2, Pyruvat und Acetyl Coenzym A.

Die Vorgehensweise für die hier vorgestellte Fluss-basierte Zentralität unterscheidet sich in zwei Punkten von den Vorgehensweisen in den oben genannten Publikationen: Einerseits in der Aufbereitung der zu untersuchenden Reaktionsnetze und andererseits in der Berechnung der Zentralitätswerte für die einzelnen Metaboliten.

Bei der Aufbereitung der Reaktionsnetze betrachtet nur Arita den tatsächlichen Ablauf der Reaktion (bzgl. der Kohlenstoffatome) im Detail, in allen anderen Publikationen wird hingegen die Teilnahme eines Metaboliten an einer Reaktion als „ausreichend“ für eine Verbindung angesehen. Allerdings nutzt Arita die Informationen über die Kohlenstoffatome nicht konsequent genug aus. Die Menge der übertragenen Atome wird bei Arita nicht betrachtet, im Unterschied zur hier vorgestellten Fluss-basierten Zentralität.

Die Berechnung der Zentralitätswerte erfolgt bei allen oben genannten Publikationen mit den bereits bekannten Zentralitäten *Degree* und *Closeness*. Dabei wird in keiner der Publikationen hinterfragt, ob diese Zentralitäten für die Analyse von Metabolitgraphen eingesetzt werden können bzw. sollten. Insbesondere eine Berücksichtigung von Flussinformationen ist in keiner der Publikationen vorgenommen worden.

Die Fluss-basierte Zentralität hingegen beachtet die beiden genannten biochemisch relevanten Aspekte. Das Resultat der Analyse, bei Betrachtung der summierten Zentralitätswerte, entspricht damit dann auch genau dem erwarteten Resultat: Die Metaboliten des Zentralstoffwechsels werden als wichtig eingestuft und Kofaktoren und Donatoren[2] tauchen nicht unten den Top-30 Metaboliten auf. Zusätzlich erlaubt die Fluss-basierte Zentralität noch die Betrachtung der Wachstumsbedingungen bei der Bildung der Reihenfolge der Metaboliten, eine Eigenschaft, die in keiner der obigen Publikationen auch nur annähernd betrachtet wird.

6.5.4. Clusterung der Metaboliten anhand der Fluss-basierten Zentralitätswerte

Für eine weitere Analyse der Zentralitätswerte wurden diese hierarchisch geclustert und als Heatmap dargestellt, siehe Abbildung 6.5 bzw. Abbildung 6.6 für einen Ausschnitt aus dem Dendrogramm. Im Folgenden werden die in der Abbildung 6.5 markierten Gruppen 1–6 bzw. A–C diskutiert und anschließend wird aus den so gewonnenen Informationen der Kern des Metabolismus und die drei wichtigsten Stoffwechselwege von *E. coli* abgeleitet.

Die Metaboliten in der Gruppe 1 zeichnen sich durch einen hohen Zentralitätswert aus,

[2]Bezeichnung für Moleküle, die innerhalb einer Reaktion eine funktionelle Gruppe abgeben.

wobei zwischen den aeroben (Gruppen A+B) und den anaeroben Wachstumsbedingungen (Gruppe C) ein deutlicher Unterschied sichtbar ist. Die Vergrößerung eines Teils des Dendrogramms (siehe Abbildung 6.6) zeigt, dass der Teilcluster mit den höchsten Zentralitätswerten im Wesentlichen aus Metaboliten des Zentralstoffwechsels (vgl. Abbildung 6.4) von *E. coli* besteht. Es wird ebenfalls deutlich, dass sich, bzgl. der Fluss-basierten Zentralität, die Metaboliten der drei bekannten Stoffwechselwege im Zentralstoffwechsel von *E. coli* jeweils sehr ähnlich verhalten. Entgegen der Erwartung gibt es allerdings folgende Unterschiede zwischen der Zuordnung der Metaboliten zu den Stoffwechselwegen und der Clusterung: Pyruvat (PYR) und Acetyl CoA (AcCoA) sind eher dem Zitronensäurezyklus als der Glykolyse zugeordnet und D-Glukose 6-phosphat (G6P), 6-Phospho-D-Glucono 1,5-Lakton (6PGL), und 6-Phospho D-Gluconat (6PGC) bilden einen eigenen Cluster, der sich von den anderen Clustern des Zentralstoffwechsels unterscheidet.

Deutlich erkennbar (Gruppe 2 und 4) sind die hohen Zentralitätswerte für die Metaboliten, die bei einer spezifischen Wachstumsbedingung zugeführt werden.

Die Gruppen 3 und 5 bestehen aus Metaboliten, die für die Produktion der Biomasse bzw. deren Vorprodukte erforderlich sind. Die Zentralitätswerte in diesen beiden Teilclustern sind jeweils relativ niedrig und variieren nicht so stark, wie die Zentralitätswerte in der Gruppe 1.

Die Metaboliten in der 6. Gruppe haben für alle Wachstumsbedingungen jeweils einen Zentralitätswert nahe an Null.

Zusätzlich zur Clusterung der Metaboliten erfolgte eine Clusterung der Wachstumsbedingungen (Gruppen A-C). Anhand der gebildeten drei Gruppen können unterschiedliche Verhaltensweisen, in Abhängigkeit von den verfügbaren Metaboliten, für den simulierten Organismus erkannt werden. Die in den Wachstumsbedingungen der Gruppe A zugeführten Metaboliten müssen alle über die Glykolyse bzw. den Pentosephosphat-Stoffwechselweg transformiert werden. Folglich haben die Metaboliten des Zitronensäurezyklus einen geringeren Zentralitätswert als die Metaboliten der Glykolyse bzw. des Pentosephosphat-Stoffwechselwegs. Die Gruppe B enthält die Wachstumsbedingungen, deren Metaboliten nicht über die Glykolyse bzw. den Pentosephosphat-Stoffwechselweg transformiert werden müssen. In diesem Fall sind die Zentralitätswerte der Metaboliten des Zitronensäurezyklus höher als die der anderen beiden Stoffwechselwege. Deutlich sichtbar (Gruppe C) wird, dass unter anaeroben Wachstumsbedingungen die Zentralitätswerte der Metaboliten signifikant niedriger sind, als unter aeroben. Dieses deckt sich mit der Beobachtung, dass *E. coli* unter anaeroben Bedingungen deutlich schlechter wächst.

Bemerkenswert an den Resultaten der Clusterung sind zwei Aspekte: einerseits, dass die Metaboliten der Gruppe 1 fast identisch zu den Metaboliten des Zentralstoffwechsels des untersuchten Organismus sind und zweitens, dass die einzelnen Teilcluster innerhalb der Gruppe 1 ziemlich genau den drei Stoffwechselwegen des Zentralstoffwechsels entsprechen.

Für den ersten Aspekt, der Identifikation der Metaboliten des Zentralstoffwechsels eines Organismus, wurden bereits Verfahren von *Ma et al.* [112] und *Almaas et al.* [7, 5] vorgestellt. *Ma et al.* nutzten dabei ein Netzwerk-Clustering-Verfahren ohne Einsatz von Flüssen [112] und *Almaas et al.* verwendeten zwar Flussinformationen, allerdings sind beispielsweise keine Metaboliten des Zitronensäurezyklus im identifizierten „*metabolic core*“ enthalten [7]. Dieses widerspricht der Idee eines „*metabolic core*“, da aus den Metaboliten des „Kerns“ möglichst alle anderen Metaboliten (indirekt) ableitbar sein sollten. Der hier vorgestellte Weg über die Fluss-basierte Zentralität fügt diesen Vorschlägen einen weiteren hinzu, wobei er Flussinformationen verwendet und so die tatsächlichen Abläufe innerhalb des Reaktionsnetzes wesentlich genauer nachbildet und ein Ergebnis erzielt, dass der Erwartung entspricht.

Auch für den zweiten Aspekt, der Detektion von Stoffwechselwegen in einem Organismus,

wurden bereits von anderen Autoren Verfahren vorgeschlagen [72, 112, 141]. In keiner der drei Veröffentlichungen wurden allerdings Flussinformationen für die Berechnung der zusammengehörigen Stoffwechselwege genutzt. In der hier vorgeschlagenen Methode entsprechen die Teilcluster aus der Gruppe 1 fast genau den drei Stoffwechselwegen des Zentralstoffwechsels von *E. coli*.

Die Fluss-basierte Zentralität erlaubt somit, zusätzlich zur Erstellung der Reihenfolge der Metaboliten, die unvoreingenommene Bestimmung der zentralen Elemente des Metabolismus eines gegebenen Organismus und die Einteilung von Metaboliten in Stoffwechselwege.

6.6. Zusammenfassung

In diesem Kapitel wurde ein Zentralitätsmaß für die Analyse von metabolischen Reaktionsnetzen vorgestellt. Wie jedes andere Zentralitätsmaß auch, ist es in der Lage, die innerhalb des Netzwerkes verwendeten Metaboliten in eine Reihenfolge zu bringen. Wesentlicher Unterschied zu allen bisher existierenden (bzw. verwendeten) Zentralitäten für die Analyse von metabolischen Reaktionsnetzen ist allerdings die Eigenschaft, dass Flussinformationen in die Bewertung der Metaboliten mit einfließen. Hierdurch wird überhaupt erst die Möglichkeit geschaffen, dass Informationen aus dem „tatsächlichen“ Ablauf innerhalb eines Organismus in die Bewertung der Metaboliten mit einfließen.

Demonstriert wurde die Leistungsfähigkeit des neuen Zentralitätsmaßes anhand eines Reaktionsnetzes für *E. coli*. Für diesen Organismus wurde zuerst eine Reihenfolge der Metaboliten bestimmt, dann wurden diese Metaboliten geclustert und anhand der Clusterung-Resultate konnte sowohl ein „*metabolic core*“ als auch die drei wichtigsten Stoffwechselwege identifiziert werden. Ein Vergleich der ermittelten Bewertungen der Metaboliten mit bereits publizierten Ergebnissen zeigt deutlich, dass ohne die Betrachtung des Flusses eine Bewertung von Metaboliten nur unzureichend möglich ist.

Das vorgestellte Zentralitätsmaß lässt sich noch auf vielfältige Weise erweitern. Nennenswert sind a. die Erweiterung der Betrachtung auf andere Atome, beispielsweise Phosphor oder Stickstoff, und b. die Einbringung von gemessenen Flusswerten, als Variation zu den hier verwendeten (über die FBA) berechneten Werten.

Datentyp	Funktionsname	Beschreibung	Laufzeit
Graph	createEmptyGraph	Erstellt einen Graphen ohne Knoten und Kanten.	$\mathcal{O}(1)$
Graph	getOrCreateVertex	Erstellt einen neuen Knoten mit dem übergebenen Namen als Knotenlabel oder gibt den zum übergebenen Label gehörenden Knoten zurück.	$\mathcal{O}(\|V\|)$
Graph	createEdge	Erstellt eine Kante zwischen den übergebenen Knoten mit dem übergebenen Kantengewicht.	$\mathcal{O}(\|V\|)$
Graph	deleteEdge	Entfernt die vom Startknoten zum Zielknoten zeigende Kante.	$\mathcal{O}(\|V\| + \|E\|)$
Graph	findEdge	Gibt die Kante zwischen den übergebenen Knoten zurück bzw. gibt Null zurück, falls keine Kante existiert.	$\mathcal{O}(\|V\| + \|E\|)$
Graph	getAntiParallelEdge	Gibt die zur übergebenen Kante entsprechende antiparallele Kante zurück. Falls keine antiparallele Kante existiert wird Null zurückgegeben.	$\mathcal{O}(\|V\| + \|E\|)$
Graph	getEdgeWeight	Gibt das Kantengewicht der übergebenen Kante zurück.	$\mathcal{O}(1)$
Graph	setEdgeWeight	Setzt das Kantengewicht der übergebenen Kante auf den übergebenen Wert.	$\mathcal{O}(1)$
Graph	getSubstrateNames	Gibt die Namen der Knoten (Metaboliten), die durch eingehende Kanten mit der durch den übergebenen Namen bezeichneten Reaktion verbunden sind zurück.	$\mathcal{O}(\|V\| + \|E\| * \|V\|)$
Graph	getProductNames	Wie getSubstrateNames allerdings für ausgehende Kanten.	$\mathcal{O}(\|V\| + \|E\| * \|V\|)$
Graph	getAllEdges	Gibt eine Liste aller Kanten zurück.	$\mathcal{O}(\|V\| + \|E\|)$
Kohlenstoffvert.	getCarbonTransfer	Gibt die Anzahl der innerhalb der übergebenen Reaktion zwischen den beiden Metaboliten übertragenen Kohlenstoffatome zurück.	$\mathcal{O}(\log_2(\|R\|) * \log_2(\|M\|)^2)$
Menge	removeElement	Entfernt das übergebene Element aus der Menge.	$\mathcal{O}(\log_2(\|S\|))$
Menge	getOneElement	Gibt ein beliebiges Element aus der Menge zurück.	$\mathcal{O}(1)$
Menge	isEmpty	Gibt Wahr zurück, falls die Menge leer ist.	$\mathcal{O}(1)$

Tabelle 6.1.: Beschreibung der in den Algorithmen 6.1 und 6.2 verwendeten Funktionen. Die in der Spalte Laufzeit verwendeten Bezeichner sind: V, Knotenmenge; E, Kantenmenge; R, Menge der Reaktionen; M, Menge der Metaboliten; S, die betrachtete Menge.

Datentyp	Beschreibung
Graph	Implementiert als Adjazenzliste mit einer Liste pro Knoten für die Nachfolger-Knoten. Jeder Knoten wird durch eine Nummer identifiziert und besitzt zusätzlich einen Knotennamen als weiteres Attribut. Die Kantengewichte werden als zusätzliche Attribute in der Adjazenzliste abgelegt.
Kohlenstoffvert.	Implementiert als mehrfach verschachtelter AVL-Baum.
Flussvektor	Implementiert als Liste mit zwei Komponenten pro Eintrag (Name & Flusswert).
Menge	Implementiert als AVL-Baum.
Liste	Implementiert als verkettete Liste.

Tabelle 6.2.: Beschreibung der in den Algorithmen 6.1 und 6.2 verwendeten Datentypen.

Reaktionsbez.	Metabolit 1	Metabolit 2	# C-Atome
Transketolase	D-Ribose-5-P	D-Sedoheptulose-7-P	5
Transketolase	Xylulose-5-P	D-Sedoheptulose-7-P	2
Transketolase	Xylulose-5-P	D-Glyceraldehyd-3-P	3

Tabelle 6.3.: Einträge in der Datenstruktur Kohlenstoffverteilung für die Reaktion Transketolase (siehe Abbildung 6.3 (rechts)).

Bezeichnung	Metabolitname (Summenformel)	Max. Aufnahme	O_2 verfügbar?
AC +	Azetat ($C_2H_3O_2$)	30,00	Ja
AKG +	α-Ketoglutarat ($C_5H_4O_5$)	12,00	Ja
ALA +	L-Alanin ($C_3H_7NO_2$)	20,00	Ja
GLC +	D-Glukose ($C_6H_{12}O_6$)	10,00	Ja
GLCN +	D-Glukonat ($C_6H_{11}O_7$)	10,00	Ja
GLYC +	Glyzerin ($C_3H_8O_3$)	20,00	Ja
LAC +	D-Laktat ($C_3H_5O_3$)	20,00	Ja
LCTS +	Laktose ($C_{12}H_{22}O_{11}$)	5,00	Ja
MAL +	L-Malat ($C_4H_4O_5$)	15,00	Ja
OCDCA +	Stearinsäure ($C_{18}H_{35}O_2$)	3,33	Ja
PRO +	L-Prolin ($C_5H_9NO_2$)	12,00	Ja
PYR +	Pyruvat ($C_3H_3O_3$)	20,00	Ja
RIB +	D-Ribose ($C_5H_{10}O_5$)	12,00	Ja
SBT +	D-Sorbitol ($C_6H_{14}O_6$)	10,00	Ja
SUCC +	Succinat ($C_4H_4O_4$)	15,00	Ja
GLC -	D-Glukose ($C_6H_{12}O_6$)	10,00	Nein
GLCN -	D-Glukonat ($C_6H_{11}O_7$)	10,00	Nein
RIB -	D-Ribose ($C_5H_{10}O_5$)	12,00	Nein
SBT -	D-Sorbitol ($C_6H_{14}O_6$)	10,00	Nein

Tabelle 6.4.: Die wesentlichen Eigenschaften der 19 verwendeten Wachstumsbedingungen. Alle Werte in der Spalte „Max. Aufnahme“ sind in der Einheit $\frac{\text{mmol}}{\text{h}}\frac{\text{g}}{\text{DW}}$ angegeben. Hierbei steht DW für das Trockengewicht (engl. *dry weight*), d.h. das Gewicht der Probe abzüglich des enthaltenen Wassers.

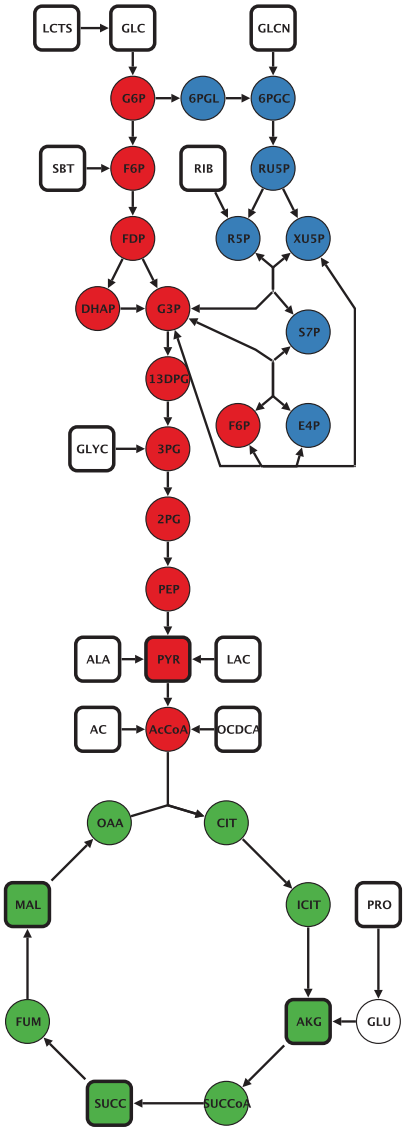


Abbildung 6.4.: Der Zentralstoffwechsel von *E. coli* als Metabolitgraph (vereinfachte Darstellung). Die in den drei Stoffwechselwegen Glykolyse (rot), Zitronensäurezyklus (grün) und Pentosephosphat-Stoffwechselweg (blau) wichtigsten Metaboliten sind jeweils farblich markiert. Zusätzlich sind die für die 19 Wachstumsbedingungen als Substratmetabolit genutzten Metaboliten als Quadrate in die Zeichnung mit aufgenommen worden. Abkürzungen der Metaboliten: siehe Tabelle 6.4 und Tabelle 6.5 und 6PGL: 6-Phospho-D-Glucono 1,5-Lakton und GLU: L-Glutamat. Quelle: [94] (Eigene Publikation), Visualisierung: VANTED [81]

Position	Metabolit-ID	Metabolitname	$\sum$ Zentralitätswerte	Stoffwechselweg
1	G3P	Glycerinaldehyd-3-phosphat	11584, 21	Glykolyse
2	F6P	D-Fruktose 6-phosphat	11564, 81	Glykolyse
3	13DPG	1,3-Bisphosphoglycerat	10705, 56	Glykolyse
4	AcCoA	Acetyl CoA	10619, 46	Glykolyse
5	3PG	3-Phospho D-Glycerat	10486, 82	Glykolyse
6	PYR	Pyruvat	10204, 73	Glykolyse
7	MAL	L-Malat	10157, 91	TCA
8	FUM	Fumarat	10109, 16	TCA
9	MALCoA	Malonyl CoA	10061, 77	
10	CoA	Coenzym A	10030, 76	
11	OAA	Oxalacetat	10009, 09	TCA
12	SUCC	Succinat	9942, 48	TCA
13	RU5P	D-Ribulose 5-phosphat	9878, 29	PPP
14	CIT	Citrat	9873, 87	TCA
15	2PG	D-Glycerat 2-phosphat	9872, 83	Glykolyse
16	ICIT	Isocitrat	9694, 15	TCA
17	DHAP	Dihydroxyacetonphosphat	9652, 38	Glykolyse
18	AKG	α-Ketoglutarat	9641, 94	TCA
19	PEP	Phosphoenolpyruvat	9547, 94	Glykolyse
20	SUCCoA	Succinyl CoA	9170, 58	TCA
21	R5P	α-D-Ribose 5-phosphat	9064, 74	PPP
22	XU5P	D-Xylulose 5-phosphat	8762, 09	PPP
23	FDP	D-Fruktose 1,6-bisphosphat	8130, 85	Glykolyse
24	G6P	D-Glukose 6-phosphat	7173, 67	Glykolyse
25	S7P	Sedoheptulose-7-phosphat	6975, 80	PPP
26	E4P	D-Erythrose 4-phosphat	6389, 81	PPP
27	ASP	L-Aspartat	6340, 43	
28	CO2	CO_2	6002, 98	
29	6PGC	6-Phospho D-Gluconat	5847, 29	PPP
30	PRPP	5-Phospho D-Ribose 1-diphosphat	5567, 11	

Tabelle 6.5.: Die Top-30 Metaboliten anhand der summierten Zentralitätswerte für die 19 Wachstumsbedingungen. Die Farben in der Spalte Stoffwechselweg sind konsistent mit der Darstellung in der Abbildung 6.4. Abkürzungen: TCA: Zitronensäurezyklus, PPP: Pentosephosphat-Stoffwechselweg. Quelle: [94] (Eigene Publikation)

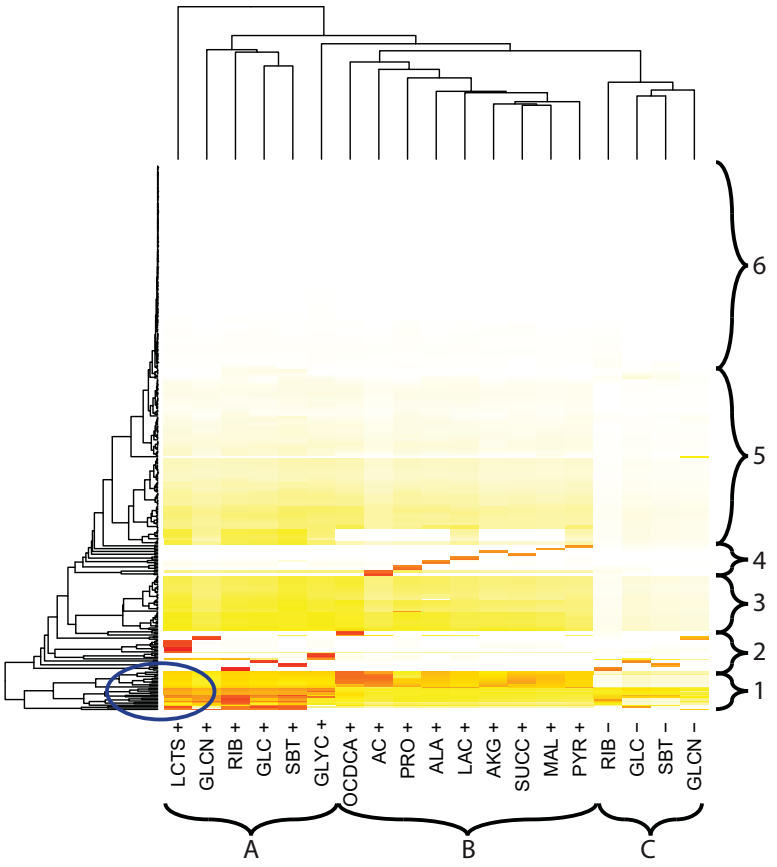


Abbildung 6.5.: Heatmap der Zentralitätswerte für die Metaboliten von *E. coli* unter den 19 Wachstumsbedingungen. Hohe Zentralitätswerte sind durch rote Balken, mittlere Werte durch gelbe Balken und niedrige Werte bzw. der Wert 0 als weiße Balken dargestellt. Die Zentralitätswerte wurden hierarchisch geclustert (Distanzmaß: euklidisch, Methode: *complete linkage*, Software: R [139]). Die Gruppen 1–6 und A–C werden im Text diskutiert, der umrandete Bereich ist in der Abbildung 6.6 vergrößert dargestellt. Die Wachstumsbedingungen sind am unteren Rand bezeichnet. Ein Pluszeichen (+) markiert aerobes und ein Minuszeichen (−) anaerobes Wachstum. Die Abkürzungen der Metaboliten sind in den Tabellen 6.4 und 6.5 angegeben. Quelle: [94] (Eigene Publikation)

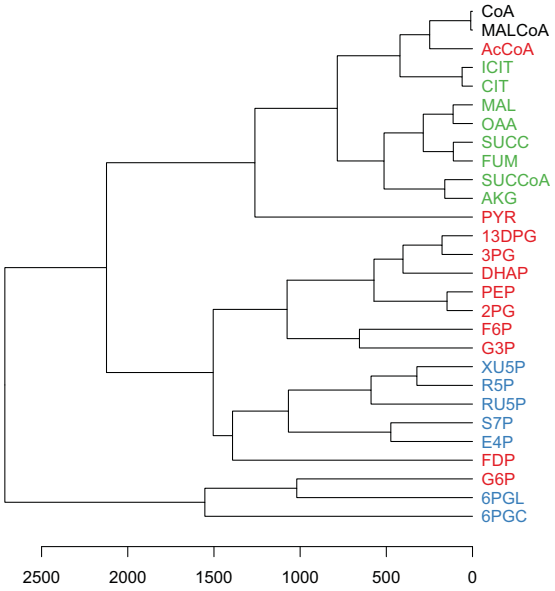


Abbildung 6.6.: Ausschnittsvergrößerung des Dendrogramms aus Abbildung 6.5. Dargestellt ist der Teilcluster mit den höchsten Zentralitätswerten. Dieser besteht im Wesentlichen aus Metaboliten der Glykolyse, des Pentosephosphat-Stoffwechselweges und des Zitronensäurezyklus. Die Farben der Metaboliten entsprechen der Darstellung in Abbildung 6.4. Abkürzungen der Metaboliten: siehe Tabelle 6.4, Tabelle 6.5 und Abbildung 6.4. Quelle: [94] (Eigene Publikation)

7. Zusammenfassung und Ausblick

Diese Dissertation schließt mit einer Zusammenfassung und einem Ausblick auf noch offene Fragestellungen im Kontext der Zentralitätsanalyse molekularbiologischer Netzwerke.

Zusammenfassung

Die Zentralitätsanalyse molekularbiologischer Netzwerke, d.h. die Bestimmung einer Reihenfolge molekularbiologisch relevanter Objekte (z.B. Transkriptionsfaktoren, Proteinen, Metaboliten), war das übergeordnete Thema dieser Dissertation. Ermittelt wurde diese Reihenfolge nicht durch Experimente oder eine Literaturrecherche, sondern allein auf der Basis der Netzwerkstruktur, in die die betrachteten Objekte eingebettet sind. Für Transkriptionsfaktoren lässt sich diese Reihenfolge beispielsweise auf der Basis eines Genregulationsnetzwerkes ermitteln und für Metaboliten auf der Basis eines metabolischen Reaktionsnetzes.

Einige Anwendungsfelder der Zentralitätsanalyse, die Rückführung der allgemeinen Idee auf die Analyse sozialer Netzwerke und eine erste Motivation für die Entwicklung neuer Zentralitätsmaße für molekularbiologische Netzwerke wurde im Kapitel 1 beschrieben. Im Kapitel 2 wurden alle notwendigen Grundlagen aus der Graphentheorie einheitlich dargestellt und bereits bekannte Zentralitätsmaße und ihre Eigenschaften dokumentiert. Die notwendigen Grundlagen aus der Molekularbiologie wurden im Kapitel 3 überblicksartig zusammengefasst. Und im Kapitel 4 wurde, basierend auf den bereits publizierten Zentralitätsanalysen, die Notwendigkeit von angepassten Zentralitätsmaßen für die Analyse molekularbiologischer Netzwerke motiviert.

Im Kapitel 5 wurde das erste der in dieser Dissertation vorgestellten Zentralitätsmaße beschrieben. Dieses Zentralitätsmaß nutzt zur Bestimmung der Reihenfolge der Transkriptionsfaktoren die Häufigkeiten des Auftretens von Netzwerkmotiven. Bei dieser Betrachtung wird davon ausgegangen, dass ein Knoten, der in vielen verschiedenen Instanzen eines gesuchten Netzwerkmotives auftritt eine besonders große Rolle innerhalb des untersuchten Netzwerkes spielt; folglich soll diesem Objekt ein höherer Zentralitätswert zugeordnet werden. Zusätzlich zur ersten Variante, bei der die Motive gleichberechtigt in die Berechnung der Zentralitätswerte einfließen, wurden zwei weitere Varianten der Zentralität vorgestellt: die Rollen-basierte Motiv-basierte Zentralitätsfamilie und die Rollen-basierte Motiv-basierte Zentralitätsfamilie für Klassen von Motiven. Alle drei Varianten zeichnen sich dadurch aus, dass sie „konfigurierbar“ sind und jeweils nur einen Teilbereich des Netzes zur Bestimmung des Zentralitätswertes einsetzen. Angewendet wurden die neu vorgestellten Zentralitätsmaße zur Analyse eines Genregulationsnetzes von *E. coli*. Dabei konnte gezeigt werden, dass die hier vorgestellten Zentralitätsmaße zur Bestimmung von globalen Regulatoren sehr gut geeignet sind und in der Qualität der ermittelten Reihenfolge die bisher existierenden Zentralitätsmaße übertreffen.

Für die Analyse metabolischer Reaktionsnetze wurde im Kapitel 6 die Fluss-basierte Zentralität vorgestellt. Kernelement dieses Zentralitätsmaßes zur Bestimmung einer Reihenfolge der Metaboliten ist die Verwendung von Informationen über den Kohlenstoffffluss. Angewendet wurde diese Zentralität wiederum auf ein Netzwerk für den Organismus *E. coli*. Bei dieser

Analyse wurde eine Reihenfolge der Metaboliten bestimmt und darauf aufbauend ein „*metabolic core*“ und die drei wichtigsten Stoffwechselwege von *E. coli* identifiziert. Wiederum zeigte ein Vergleich mit Resultaten auf der Basis der bereits bekannten Zentralitäten die Überlegenheit der hier neu vorgestellten Zentralität.

Für beide hier vorgestellte Zentralitätsmaße gilt, dass diese auf neuen Überlegungen basieren. Bei den Motiv-basierten Zentralitäten besteht die Neuerung in der Verwendung von Teilgraphen zur Berechnung der Zentralitätswerte und bei der Fluss-basierten Zentralität ist die Kombination aus Informationen über die Graphstruktur und Flussinformationen vorher noch nicht publiziert worden. Beide Zentralitätsmaße zeichnen sich dadurch aus, dass sie die bereits existierenden Zentralitätsmaße in ihrer Leistungsfähigkeit deutlich übertreffen.

Ausblick

Die hier vorgestellten Zentralitätsmaße lassen sich noch auf vielfältige Weise verbessern bzw. optimieren.

Bei der Motiv-basierten Zentralität sind Verbesserungen in Hinblick auf die algorithmische Umsetzung die wichtigste zu lösende Aufgabe. Aufgrund der dem Problem inhärenten Komplexität ist dabei sicherlich auch eine approximative Lösung denkbar. Eine mögliche Erweiterung besteht in der Betrachtung von Netzwerken mit zusätzlichen Eigenschaften. Hierbei sind Kantengewichte die naheliegende Information. Aber auch die Verwendung von gefärbten Knoten (und Kanten), sowohl für den Quellgraph als auch für das Motiv, sind vielversprechende Ansätze.

Für die Fluss-basierte Zentralität sind Erweiterungen in zwei Richtungen denkbar. Aus der Sicht der Biologie dürfte dabei die Einbringung von gemessenen Flusswerten, im Unterschied zu den auf der Basis der FBA berechneten, die interessantere Erweiterung sein. Eine Anpassung an die Betrachtung anderer Atome, beispielsweise Phosphor oder Stickstoff, ist auf den ersten Blick sicherlich trivial, da am Verfahren an sich nicht geändert werden muss. Allerdings könnte diese Fragestellung dennoch interessante Einblicke in den Metabolismus gewähren. Und spätestens bei der gemeinsamen Betrachtung mehrerer Atome und, für die Zukunft sicherlich zu erwartender, größerer Netze wird auch die Fragestellung nach verbesserten Algorithmen für die Berechnung der Zentralitätswerte relevant.

Literaturverzeichnis

[1] Ahuja, R. K. ; Magnanti, T. L. ; Orlin, J. B.: *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993

[2] Aittokallio, T. ; Schwikowski, B. : Graph-based methods for analysing networks in cell biology. In: *Briefings in Bioinformatics* 7 (2006), Nr. 3, S. 243–255

[3] Albert, R. : Scale-free networks in cell biology. In: *Journal of Cell Science* 118 (2005), Nr. 21, S. 4947–4957

[4] Alberts, B. ; Bray, D. ; Johnson, A. ; Lewis, J. ; Raff, M. ; Roberts, K. ; Walter, P. : *Lehrbuch der Molekularen Zellbiologie*. Wiley-VCH, 1999

[5] Almaas, E. ; Kovács, B. ; Vicsek, T. ; Oltvai, Z. N. ; Barabási, A.-L. : Global organization of metabolic fluxes in the bacterium Escherichia coli. In: *Nature* 427 (2004), Nr. 6977, S. 839–843

[6] Almaas, E. : Biological impacts and context of network theory. In: *Journal of Experimental Biology* 210 (2007), S. 1548–1558

[7] Almaas, E. ; Oltvai, Z. N. ; Barabási, A.-L. : The Activity Reaction Core and Plasticity of Metabolic Networks. In: *PLoS Computational Biology* 1 (2005), Nr. 7, S. e68

[8] Alon, N. ; Yuster, R. ; Zwick, U. : Finding and Counting Given Length Cycles. In: *Algorithmica* 17 (1997), Nr. 3, S. 209–223

[9] Alon, U. : Network motifs: theory and experimental approaches. In: *Nature Reviews Genetics* 8 (2007), Nr. 6, S. 450–461

[10] Aloy, P. ; Russell, R. B.: Potential artefacts in protein-interaction networks. In: *FEBS Lett* 530 (2002), Nr. 1-3, S. 253–254

[11] Anthonisse, J. M.: The rush in a directed graph / Stichting Mathematisch Centrum. 1971 (BN 9/71). – Forschungsbericht

[12] Anton, H. ; Rorres, C. : *Elementary Linear Algebra: Applications Version*. 7. John Wiley & Sons, 1994

[13] Arikati, S. R. ; Chaudhuri, S. ; Zaroliagis, C. D.: All-Pairs Min-Cut in Sparse Networks. In: *Journal of Algorithms* 29 (1998), Nr. 1, S. 82–110

[14] Arita, M. : The metabolic world of *Escherichia coli* is not small. In: *Proc Natl Acad Sci U S A* 101 (2004), Nr. 6, S. 1543–1547

[15] Bader, G. D. ; Betel, D. ; Hogue, C. W. V.: BIND: the Biomolecular Interaction Network Database. In: *Nucleic Acids Research* 31 (2003), Nr. 1, S. 248–250

[16] Bader, G. D. ; Cary, M. P. ; Sander, C. : Pathguide: a Pathway Resource List. In: *Nucleic Acids Research* 34 (2006), Nr. Database-Issue, S. 504–506

[17] Barabási, A.-L. ; Oltvai, Z. N.: Network biology: understanding the cell's functional organization. In: *Nature Reviews Genetics* 5 (2004), Nr. 2, S. 101–113

[18] Batada, N. N. ; Hurst, L. D. ; Tyers, M. : Evolutionary and Physiological Importance of Hub Proteins. In: *PLoS Computational Biology* 2 (2006), Nr. 7, S. e88

[19] Becker, S. A. ; Feist, A. M. ; Mo, M. L. ; Hannum, G. ; Palsson, B. Ø. ; Herrgard, M. J.: Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. In: *Nature Protocols* 2 (2007), S. 727–738

[20] Berg, J. M. ; Stryer, L. ; Tymoczko, J. L.: *Biochemie*. Spektrum Akademischer Verlag, 2007

[21] Bonacich, P. : Factoring and Weighting Approaches to Status Scores and Clique Identification. In: *Journal of Mathematical Sociology* 2 (1972), S. 113–120

[22] Borgatti, S. P.: Centrality and Network Flow. In: *Social Networks* 27 (2005), S. 55–71

[23] Borgatti, S. P. ; Everett, M. G.: A Graph-theoretic perspective on centrality. In: *Social Networks* 28 (2006), S. 466–484

[24] Brandes, U. ; Erlebach, T. : Fundamentals. In: *N*etwork Analysis: Methodological Foundations[25], S. 7–15

[25] Brandes, U. (Hrsg.) ; Erlebach, T. (Hrsg.): *LNCS Tutorial*. Bd. 3418: *Network Analysis: Methodological Foundations*. Springer, 2005

[26] Brandes, U. ; Fleischer, D. : Centrality Measures Based on Current Flow. In: *Proc. 22nd Symp. Theoretical Aspects of Computer Science (STACS '05)* Bd. 3404, Springer-Verlag, 2005 (Lecture Notes in Computer Science (LNCS)), S. 533–544

[27] Bryant, R. E.: Symbolic Boolean manipulation with ordered binary-decision diagrams. In: *ACM Comput. Surv.* 24 (1992), Nr. 3, S. 293–318

[28] Chiba, N. ; Nishizeki, T. : Arboricity and subgraph listing algorithms. In: *SIAM J. Comput.* 14 (1985), Nr. 1, S. 210–223

[29] Clark, D. P.: *Molecular Biology, Understanding the Genetic Revolution: Das Original mit Übersetzungshilfen*. Spektrum Akademischer Verlag, 2006

[30] Cook, S. A.: The Complexity of Theorem-Proving Procedures. In: *Conference Record of Third Annual ACM Symposium on Theory of Computing*, ACM, 1971, S. 151–158

[31] Cordella, L. P. ; Foggia, P. ; Sansone, C. ; Vento, M. : Performance Evaluation of the VF Graph Matching Algorithm. In: *10th International Conference on Image Analysis and Processing*, 1999, S. 1172–1177

[32] Cordella, L. P. ; Foggia, P. ; Sansone, C. ; Vento, M. : An Improved Algorithm for Matching Large Graphs. In: *International Workshop on Graph-based Representation in Pattern Recognition*, 2001, S. 149–159

[33] Cordella, L. P. ; Foggia, P. ; Sansone, C. ; Vento, M. : A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004), Nr. 10, S. 1367–1372

[34] Cortadella, J. ; Valiente, G. : A Relational View of Subgraph Isomorphism. In: Desharnais, J. (Hrsg.): *RelMiCS*, 2000, S. 45–54

[35] Costenbader, E. ; Valente, T. W.: The stability of centrality measures when networks are sampled. In: *Social Networks* 25 (2003), Nr. 4, S. 283–307

[36] Coulomb, S. ; Bauer, M. ; Bernard, D. ; Marsolier-Kergoat, M.-C. : Gene essentiality and the topology of protein interaction networks. In: *Proceedings of the Royal Society B: Biological Sciences* 272 (2005), Nr. 1573, S. 1721–1725

[37] Demetrius, L. ; Manke, T. : Robustness and network evolution – an entropic principle. In: *Physica A* 346 (2005), S. 682–696

[38] Diestel, R. : *Graphentheorie*. 2. Springer-Verlag, 2000

[39] Domschke, W. ; Drexl, A. : *Location and Layout Planning: An International Bibliography*. Berlin : Springer-Verlag, 1985

[40] Doyle, P. G. ; Snell, J. L.: *Random Walks and Electric Networks*. Dokument unter der GNU General Public License, Version 3.02, arXiv:math.PR/0001057, 2000

[41] Edwards, J. S. ; Palsson, B. O.: The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. In: *Proc Natl Acad Sci U S A* 97 (2000), Nr. 10, S. 5528–5533

[42] Engelfriet, J. ; Rozenberg, G. : Node Replacement Graph Grammars. In: Rozenberg, G. (Hrsg.): *Handbook of Graph Grammars and Computing by Graph Transformation* Bd. 1. World Scientific, 1997, Kapitel 1, S. 1–94

[43] Eppstein, D. : Finding the k-Shortest Paths. In: *SIAM Journal on Computing* 28 (1999), Nr. 2, S. 652–673

[44] Eppstein, D. : Subgraph Isomorphism in Planar Graphs and Related Problems. In: *Journal of Graph Algorithms and Applications* 3 (1999), Nr. 3, S. 1–27

[45] Estrada, E. : Protein bipartivity and essentiality in theyeast protein-protein interaction network. In: *Journal of Proteome Research* 5 (2006), Nr. 9, S. 2177–2184. – Comparative Study

[46] Estrada, E. : Virtual identification of essential proteins within the protein interaction network of yeast. In: *Proteomics* 6 (2006), Nr. 1, S. 35–40

[47] Estrada, E. ; Rodríguez-Velázquez, J. A.: Spectral measures of bipartivity in complexnetworks. In: *Physical Review E* 72 (2005), S. 046105

[48] Estrada, E. ; Rodríguez-Velázquez, J. A.: Subgraph centrality in complex networks. In: *Physical Review E* 71 (2005), Nr. 056103

[49] Fell, D. A. ; Wagner, A. : The small world of metabolism. In: *Nature Biotechnology* 18 (2000), S. 1121–1122

[50] Fell, D. A. ; Wagner, A. : Structural properties of metabolic networks: implications for evolution and modelling of metabolism. In: Hofmeyr, J.-H. S. (Hrsg.) ; Rohwer, J. M. (Hrsg.) ; Snoep, J. L. (Hrsg.): *Animating the cellular map*, Stellenbosch University Press, 2000, S. 79–85

[51] Ferrar, W. L.: *Finite Matrices*. Oxford : Clarendon Press, 1951

[52] Flum, J. ; Grohe, M. : *Parameterized Complexity Theory*. Springer, 2006 (Texts in Theoretical Computer Science)

[53] Foggia, P. ; Sansone, C. ; Vento, M. : A Database of Graphs for Isomorphism and Sub-Graph Isomorphism Benchmarking. In: *International Workshop on Graph-based Representation in Pattern Recognition*, 2001, S. 176–187

[54] Foggia, P. ; Sansone, C. ; Vento, M. : A Performance Comparison of Five Algorithms for Graph Isomorphism. In: *International Workshop on Graph-based Representation in Pattern Recognition*, 2001, S. 188–199

[55] Ford, L. R. Jr. ; Fulkerson, D. R.: Maximal flow through a network. In: *Canadian Journal of Mathematics* 8 (1956), S. 399–404

[56] Fraser, H. B. ; Hirsh, A. E. ; Steinmetz, L. M. ; Scharfe, C. ; Feldman, M. W.: Evolutionary rate in the protein interaction network. In: *Science* 296 (2002), Nr. 5568, S. 750–752

[57] Freeman, L. C.: A Set of Measures of Centrality Based Upon Betweenness. In: *Sociometry* 40 (1977), S. 35–41

[58] Futschik, M. E. ; Chaurasia, G. ; Wanker, E. E. ; Herzel, H. : Comparison of Human Protein-Protein Interaction Maps. In: Huson, D. H. (Hrsg.) ; Kohlbacher, O. (Hrsg.) ; Lupas, A. N. (Hrsg.) ; Nieselt, K. (Hrsg.) ; Zell, A. (Hrsg.): *German Conference on Bioinformatics* Bd. 83, GI, 2006 (LNI). – ISBN 978–3–88579–177–5, S. 21–32

[59] Gansner, E. R. ; Koutsofios, E. ; North, S. C. ; Vo, K.-P. : A Technique for Drawing Directed Graphs. In: *IEEE Transactions on Software Engineering* 19 (1993), Nr. 3, S. 214–230

[60] Gansner, E. R. ; North, S. C.: An open graph visualization system and its applications to software engineering. In: *Software — Practice and Experience* 30 (2000), Nr. 11, 1203–1233. `citeseer.ist.psu.edu/gansner99open.html`

[61] Garey, M. R. ; Johnson, D. S.: *Computers and Intractability*. New York : W. H. Freeman and Company, 2003

[62] Ghim, C.-M. ; Goh, K.-I. ; Kahng, B. : Lethality and synthetic lethality in the genome-wide metabolic network of Escherichia coli. In: *Journal of Theoretical Biology* 237 (2005), Nr. 4, S. 401–411

[63] Godsil, C. ; Royle, G. : *Graduate Texts in Mathematics*. Bd. 207: *Algebraic Graph Theory*. Springer, 2001

[64] Goldberg, A. V. ; Tarjan, R. E.: A new approach to the maximum-flow problem. In: *Journal of the ACM* 35 (1988), Nr. 4, S. 921–940

[65] Goto, S. ; Okuno, Y. ; Hattori, M. ; Nishioka, T. ; Kanehisa, M. : LIGAND: database of chemical compounds and reactions in biological pathways. In: *Nucleic Acids Research* 30 (2002), Nr. 1, S. 402–404

[66] Grafahrend-Belau, E. ; Weise, S. ; Koschützki, D. ; Scholz, U. ; Junker, B. H. ; Schreiber, F. : MetaCrop: a detailed database of crop plant metabolism. In: *Nucleic Acids Research* 36 (2008), S. D954–958

[67] Hage, P. ; Harary, F. : Eccentricity and centrality in networks. In: *Social Networks* 17 (1995), S. 57–63

[68] Hahn, M. W. ; Conant, G. C. ; Wagner, A. : Evolution in large genetic networks: does connectivity equal constraint? In: *Journal of Molecular Evolution* 58 (2004), Nr. 2, S. 203–211

[69] Hahn, M. W. ; Kern, A. D.: Comparative genomics of centrality andessentiality in three eukaryotic protein-interaction networks. In: *Molecular Biology and Evolution* 22 (2005), Nr. 4, S. 803–806

[70] Han, J.-D. J. ; Bertin, N. ; Hao, T. ; Goldberg, D. S. ; Berriz, G. F. ; Zhang, L. V. ; Dupuy, D. ; Walhout, A. J. M. ; Cusick, M. E. ; Roth, F. P. ; Vidal, M. : Evidence for dynamically organized modularity in the yeast protein-protein interaction network. In: *Nature* 430 (2004), Nr. 6995, S. 88–93

[71] He, X. ; Zhang, J. : Why Do Hubs Tend to Be Essential in Protein Networks? In: *PLoS Genetics* 2 (2006), Nr. 6, S. e88

[72] Holme, P. ; Huss, M. ; Jeong, H. : Subnetwork hierarchies of biochemical pathways. In: *Bioinformatics* 19 (2003), Nr. 4, S. 532–538

[73] Ichikawa, S. ; Saito, H. ; Udorn, L. ; Konishi, K. : Evaluation of Accelerator Designs for Subgraph Isomorphism Problem. In: *Proceedings of 10th Int'l Conf. on Field Programmable Logic and Applications (FPL 2000)* Bd. 1896, Springer, 2000 (Lecture Notes in Computer Science). – ISBN 3–540–67899–9, S. 729–738

[74] Institute, N. H. G. R.: *Talking Glossary of Genetic Terms*. http://www.genome.gov/ glossary.cfm, 2007

[75] Jacob, R. ; Koschützki, D. ; Lehmann, K. A. ; Peeters, L. ; Tenfelde-Podehl, D. : Algorithms for Centrality Indices. In: [25], S. 62–82

[76] Jansen, R. ; Gerstein, M. : Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. In: *Current Opinion in Microbiology* 7 (2004), Nr. 5, S. 535–545

[77] Jeong, H. ; Mason, S. P. ; Barabási, A.-L. ; Oltvai, Z. N.: Lethality and centrality in protein networks. In: *Nature* 411 (2001), S. 41–42

[78] Jeong, H. ; Oltvai, Z. N. ; Barabási, A.-L. : Prediction of Protein Essentiality Based on Genomic Data. In: *Complexus* 1 (2003), S. 19–28

[79] Jeong, H. ; Tombor, B. ; Albert, R. ; Oltvai, Z. N. ; Barabási, A.-L. : The large-scale organization of metabolic networks. In: *Nature* 407 (2000), S. 651–654

[80] Joy, M. P. ; Brock, A. ; Ingber, D. E. ; Huang, S. : High-betweenness proteins in the yeast protein interaction network. In: *Journal of Biomedicine and Biotechnology* 2 (2005), S. 96–103

[81] Junker, B. H. ; Klukas, C. ; Schreiber, F. : VANTED: A system for advanced data analysis and visualization in the context of biological networks. In: *BMC Bioinformatics* 7 (2006), Nr. 109

[82] Kamada, T. ; Kawai, S. : An algorithm for drawing general undirected graphs. In: *Information Processing Letters* 31 (1989), Nr. 1, S. 7–15

[83] Kanehisa, M. : *Post-genome Informatics*. Oxford University Press, 2000

[84] Kanehisa, M. ; Goto, S. ; Hattori, M. ; Aoki-Kinoshita, K. F. ; Itoh, M. ; Kawashima, S. ; Katayama, T. ; Araki, M. ; Hirakawa, M. : From genomics to chemical genomics: new developments in KEGG. In: *Nucleic Acids Research* 34 (2006), S. D354–357

[85] Kashtan, N. ; Itzkovitz, S. ; Milo, R. ; Alon, U. : Topological generalizations of network motifs. In: *Physical Review E* 70 (2004), Nr. 031909

[86] Kashtan, N. ; Itzkovitz, S. ; Milo, R. ; Alon, U. : *mfinder Tool Guide Version 1.2*. Rehovot, Israel: Departments of Molecular Cell Biology and Computer Science & Applied Mathematics, Weizmann Institute of Science, 2005

[87] Katz, L. : A new status index derived from sociometric analysis. In: *Psychometrika* 18 (1953), Nr. 1, S. 39–43

[88] Kel-Margoulis, O. ; Matys, V. ; Choi, C. ; Reuter, I. ; Krull, M. ; Potapov, A. ; Voss, N. ; Liebich, I. ; Kel, A. ; Wingender, E. : Databases on Gene Regulation. In: Bajic, V. B. (Hrsg.) ; Wee, T. T. (Hrsg.): *Information Processing and Living Systems* Bd. 2. London : Imperial College Press, 2005, Kapitel 11, S. 709–727

[89] Kerrien, S. ; Alam-Faruque, Y. ; Aranda, B. ; Bancarz, I. ; Bridge, A. ; Derow, C. ; Dimmer, E. ; Feuermann, M. ; Friedrichsen, A. ; Huntley, R. ; Kohler, C. ; Khadake, J. ; Leroy, C. ; Liban, A. ; Lieftink, C. ; Montecchi-Palazzi, L. ; Orchard, S. ; Risse, J. ; Robbe, K. ; Roechert, B. ; Thorneycroft, D. ; Zhang, Y. ; Apweiler, R. ; Hermjakob, H. : IntAct–open source resource for molecular interaction data. In: *Nucleic Acids Research* 35 (2007), S. D561–D565

[90] Keseler, I. M. ; Collado-Vides, J. ; Gama-Castro, S. ; Ingraham, J. ; Paley, S. ; Paulsen, I. T. ; Peralta-Gil, M. ; Karp, P. D.: EcoCyc: a comprehensive database resource for *Escherichia coli*. In: *Nucleic Acids Research* 33 (2005), S. D334–337. http://dx.doi.org/10.1093/nar/gki108. – DOI 10.1093/nar/gki108

[91] Klipp, E. ; Herwig, R. ; Kowald, A. ; Wierling, C. ; Lehrach, H. : *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH, 2005

[92] Knuth, D. E.: *The Art of Computer Programming*. Bd. Volume 1, Fundamental Algorithms. Third. Addison-Wesley, 1997

[93] Köhler, W. ; Schachtel, G. ; Voleske, P. : *Biostatistik*. 3. Springer-Verlag, 2002

[94] Koschützki, D. ; Junker, B. H. ; Schwender, J. ; Schreiber, F. : Structural Analysis of Metabolic Networks based on Flux Centrality. In: *Journal of Theoretical Biology* 265 (2010), Nr. 3, S. 261–269

[95] Koschützki, D. ; Lehmann, K. A. ; Peeters, L. ; Richter, S. ; Tenfelde-Podehl, D. ; Zlotowski, O. : Centrality Indices. In: [25], S. 16–61

[96] Koschützki, D. ; Lehmann, K. A. ; Tenfelde-Podehl, D. ; Zlotowski, O. : Advanced Centrality Concepts. In: [25], S. 83–111

[97] Koschützki, D. ; Schreiber, F. : Comparison of Centralities for Biological Networks. In: *Proceedings of the German Conference on Bioinformatics 2004* Bd. P-53, Springer, 2004 (Lecture Notes in Informatics), S. 199–206

[98] Koschützki, D. ; Schreiber, F. : Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. In: *Gene Regulation and Systems Biology* 2 (2008), S. 193–201

[99] Koschützki, D. ; Schwöbbermeyer, H. ; Schreiber, F. : Ranking of network elements based on functional substructures. In: *Journal of Theoretical Biology* 248 (2007), Nr. 3, S. 471–479

[100] Krull, M. ; Pistor, S. ; Voss, N. ; Kel, A. ; Reuter, I. ; Kronenberg, D. ; Michael, H. ; Schwarzer, K. ; Potapov, A. ; Choi, C. ; Kel-Margoulis, O. ; Wingender, E. : TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. In: *Nucleic Acids Res* 34 (2006), Nr. Database issue, S. 546–551

[101] Kulikova, T. ; Akhtar, R. ; Aldebert, P. ; Althorpe, N. ; Andersson, M. ; Baldwin, A. ; Bates, K. ; Bhattacharyya, S. ; Bower, L. ; Browne, P. ; Castro, M. ; Cochrane, G. ; Duggan, K. ; Eberhardt, R. ; Faruque, N. ; Hoad, G. ; Kanz, C. ; Lee, C. ; Leinonen, R. ; Lin, Q. ; Lombard, V. ; Lopez, R. ; Lorenc, D. ; McWilliam, H. ; Mukherjee, G. ; Nardone, F. ; Pilar Garcia Pastor, M. ; Plaister, S. ; Sobhany, S. ; Stoehr, P. ; Vaughan, R. ; Wu, D. ; Zhu, W. ; Apweiler, R. : EMBL Nucleotide Sequence Database in 2006. In: *Nucleic Acids Research* 35 (2006), S. D16–D20

[102] Langville, A. N. ; Meyer, C. D.: A Survey of Eigenvector Methods for Web Information Retrieval. In: *SIAM Review* 47 (2005), Nr. 1, S. 135–161

[103] Lee, T. I. ; Rinaldi, N. J. ; Robert, F. ; Odom, D. T. ; Bar-Joseph, Z. ; Gerber, G. K. ; Hannett, N. M. ; Harbison, C. T. ; Thompson, C. M. ; Simon, I. ; Zeitlinger, J. ; Jennings, E. G. ; Murray, H. L. ; Gordon, D. B. ; Ren, B. ; Wyrick, J. J. ; Tagne, J.-B. ; Volkert, T. L. ; Fraenkel, E. ; Gifford, D. K. ; Young, R. A.: Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. In: *Science* 298 (2002), Nr. 5594, S. 799–804

[104] Lemke, N. ; Herédia, F. ; Barcellos, C. K. ; Reis, A. N. ; Mombach, J. C. M.: Essentiality and damage in metabolic networks. In: *Bioinformatics* 20 (2004), Nr. 1, S. 115–119

[105] Liu, W. ; Li, D. ; Zhang, J. ; Zhu, Y. ; He, F. : SigFlux: a novel network feature to evaluate the importance of proteins in signal transduction networks. In: *BMC Bioinformatics* 7 (2006), S. 515

[106] Liu, W. chung ; Lin, W. hsien ; Davis, A. J. ; Jordán, F. ; Yang, H. te ; Hwang, M. jing: A network perspective on the topological importance of enzymes and their phylogenetic conservation. In: *BMC Bioinformatics* 8 (2007), S. 121

[107] Lodish, H. ; Berk, A. ; Matsudaira, P. ; Kaiser, C. A. ; Krieger, M. ; Scott, M. P. ; Zipursky, L. ; Darnell, J. : *Molecular Cell Biology*. 5. W. H. Freeman and Company, 2004

[108] Lougee-Heimer, R. : The Common Optimization INterface for Operations Research. In: *IBM Journal of Research and Development* 47 (2003), Nr. 1, S. 57–66

[109] Lu, C. ; Zhang, Z. ; Leach, L. ; Kearsey, M. ; Luo, Z. : Impacts of yeast metabolic network structure on enzyme evolution. In: *Genome Biology* 8 (2007), Nr. 8, S. 407

[110] Ma, H.-W. ; Zeng, A.-P. : The connectivity structure, giant strong component and centrality of metabolic networks. In: *Bioinformatics* 19 (2003), Nr. 11, S. 1423–1430

[111] Ma, H.-W. ; Zeng, A.-P. : Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. In: *Bioinformatics* 19 (2003), Nr. 2, S. 270–277

[112] Ma, H.-W. ; Zhao, X.-M. ; Yuan, Y.-J. ; Zeng, A.-P. : Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. In: *Bioinformatics* 20 (2004), Nr. 12, S. 1870–1876

[113] MacIsaac, K. D. ; Fraenkel, E. : Practical strategies for discovering regulatory DNA sequence motifs. In: *PLoS Computational Biology* 2 (2006), Nr. 4, S. e36

[114] Mahadevan, R. ; Palsson, B. O.: Properties of metabolic networks: structure versus function. In: *Biophysical Journal* 88 (2005), Nr. 1, S. L07–L09

[115] Mahadevan, R. ; Schilling, C. : Effects of Alternate Optimal Solutions in Constraint-based Genome Scale Metabolic Models. In: *Metabolic Engineering* 6 (2003), Nr. 4, S. 264–276

[116] Mangan, S. ; Alon, U. : Structure and Function of the Feed-Forward Loop Network Motif. In: *PNAS* 100 (2003), Nr. 21, S. 11980–11985

[117] Mangan, S. ; Zaslaver, A. ; Alon, U. : The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks. In: *Journal of Molecular Biology* 334 (2003), Nr. 2, S. 197–204

[118] Manke, T. ; Demetrius, L. ; Vingron, M. : Lethality and entropy of protein interaction networks. In: *International Conference on Genome Informatics* 16 (2005), Nr. 1, S. 159–163

[119] Manke, T. ; Demetrius, L. ; Vingron, M. : An entropic characterization of protein interaction networks and cellular robustness. In: *Journal of The Royal Society Interface* 22 (2006), S. 843–850

[120] Martínez-Antonio, A. ; Collado-Vides, J. : Identifying global regulators in transcriptional regulatory networks in bacteria. In: *Current Opinion in Microbiology* 6 (2003), Nr. 5, S. 482–489

[121] Mason, O. ; Verwoerd, M. : Graph theory and networks in Biology. In: *IET Systems Biology* 1 (2007), Nr. 2, S. 89–119

[122] Matys, V. ; Kel-Margoulis, O. V. ; Fricke, E. ; Liebich, I. ; Land, S. ; Barre-Dirrie, A. ; Reuter, I. ; Chekmenev, D. ; Krull, M. ; Hornischer, K. ; Voss, N. ; Stegmaier, P. ; Lewicki-Potapov, B. ; Saxel, H. ; Kel, A. E. ; Wingender, E. : TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. In: *Nucleic Acids Res* 34 (2006), Nr. Database issue, S. 108–110

[123] Mering, C. von ; Krause, R. ; Snel, B. ; Cornell, M. ; Oliver, S. G. ; Fields, S. ; Bork, P. : Comparative assessment of large-scale data sets of protein-protein interactions. In: *Nature* 417 (2002), Nr. 6887, S. 399–403

[124] Michal, G. : *Biochemical pathways.* Spektrum Akademischer Verlag, 1999

[125] Milo, R. ; Shen-Orr, S. ; Itzkovitz, S. ; Kashtan, N. ; Chklovskii, D. ; Alon, U. : Network Motifs: Simple Building Blocks of Complex Networks. In: *Science* 298 (2002), Nr. 5594, S. 824–827

[126] Nešetřil, J. ; Poljak, S. : On the complexity of the subgraph problem. In: *Commentationes Mathematicae Universitatis Carolinae* 26 (1985), Nr. 2, S. 415–419

[127] Newman, M. E. J.: A measure of betweenness centrality based on random walks. In: *Social Networks* 27 (2005), S. 39–54

[128] Oh, M. ; Yamada, T. ; Hattori, M. ; Goto, S. ; Kanehisa, M. : Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. In: *Journal of Chemical Information and Modeling* 47 (2007), Nr. 4, S. 1702–1712

[129] Overbeek, R. ; Larsen, N. ; Pusch, G. D. ; D'Souza, M. ; Jr, E. S. ; Kyrpides, N. ; Fonstein, M. ; Maltsev, N. ; Selkov, E. : WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. In: *Nucleic Acids Research* 28 (2000), Nr. 1, S. 123–125

[130] Ozier, O. ; Amin, N. ; Ideker, T. : Global architecture of genetic interactions on the protein network. In: *Nature Biotechnology* 21 (2003), Nr. 5, S. 490–491

[131] Page, L. ; Brin, S. ; Motwani, R. ; Winograd, T. : The PageRank Citation Ranking: Bringing Order to the Web / Stanford Digital Library Technologies Project. 1998. – Forschungsbericht

[132] Palaniswamy, S. K. ; James, S. ; Sun, H. ; Lamb, R. S. ; Davuluri, R. V. ; Grotewold, E. : AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks. In: *Plant Physiology* 140 (2006), Nr. 3, S. 818–829. `http://dx.doi.org/10.1104/pp.105.072280`. – DOI 10.1104/pp.105.072280

[133] Palsson, B. O.: *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 2006

[134] Pereira-Leal, J. B. ; Audit, B. ; Peregrin-Alvarez, J. M. ; Ouzounis, C. A.: An exponential core in the heart of the yeast protein interaction network. In: *Molecular Biology and Evolution* 22 (2005), Nr. 3, S. 421–425

[135] Potapov, A. P. ; Voss, N. ; Sasse, N. ; Wingender, E. : Topology of Mammalian Transcription Networks. In: *Genome Informatics* 16 (2005), Nr. 2, S. 270–278

[136] Proulx, S. R. ; Promislow, D. E. L. ; Phillips, P. C.: Network thinking in ecology and evolution. In: *TRENDS in Ecology and Evolution* 20 (2005), Nr. 6, S. 345–353

[137] Przulj, N. ; Wigle, D. A. ; Jurisica, I. : Functional topology in a network of protein interactions. In: *Bioinformatics* 20 (2004), Nr. 3, S. 340–348. – Evaluation Studies

[138] Qi, Y. ; Ge, H. : Modularity and dynamics of cellular networks. In: *PLoS Computational Biology* 2 (2006), Nr. 12, S. e174

[139] R Development Core Team: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2005. `http://www.R-project.org`. – ISBN 3-900051-07-0

[140] Rahman, S. A. ; Schomburg, D. : Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. In: *Bioinformatics* 22 (2006), Nr. 14, S. 1767–1774

[141] Ravasz, E. ; Somera, A. ; Mongru, D. ; Oltvai, Z. N. ; Barabási, A. : Hierarchical Organization of Modularity in Metabolic Networks. In: *Science* 297 (2002), S. 1551–1555

[142] Reed, J. L. ; Vo, T. D. ; Schilling, C. H. ; Palsson, B. O.: An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). In: *Genome Biology* 4 (2003), Nr. 9, S. R54

[143] Reguly, T. ; Breitkreutz, A. ; Boucher, L. ; Breitkreutz, B.-J. ; Hon, G. ; Myers, C. ; Parsons, A. ; Friesen, H. ; Oughtred, R. ; Tong, A. ; Stark, C. ; Ho, Y. ; Botstein, D. ; Andrews, B. ; Boone, C. ; Troyanskya, O. ; Ideker, T. ; Dolinski, K. ; Batada, N. ; Tyers, M. : Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. In: *Journal of Biology* 5 (2006), Nr. 4, S. 11

[144] Rodriguez-Caso, C. ; Medina, M. A. ; Sole, R. V.: Topology, tinkering and evolution of the human transcription factor network. In: *The FEBS Journal* 272 (2005), Nr. 24, S. 6423–6434

[145] Rojas, I. ; Golebiewski, M. ; Kania, R. ; Krebs, O. ; Mir, S. ; Weidemann, A. ; Wittig, U. : SABIO-RK: a database for biochemical reactions and their kinetics. In: *BMC Systems Biology* 1 (2007), Nr. Suppl 1, S. S6

[146] Sabidussi, G. : The centrality index of a graph. In: *Psychometrika* 31 (1966), S. 581–603

[147] Salgado, H. ; Gama-Castro, S. ; Peralta-Gil, M. ; Díaz-Peredo, E. ; Sánchez-Solano, F. ; Santos-Zavaleta, A. ; Martínez-Flores, I. ; Jiménez-Jacinto, V. ; Bonavides-Martínez, C. ; Segura-Salazar, J. ; Martínez-Antonio, A. ; Collado-Vides, J. : RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. In: *Nucleic Acids Research* 34 (2006), S. D394–397. `http://dx.doi.org/10.1093/nar/gkj156`. – DOI 10.1093/nar/gkj156

[148] Salwinski, L. ; Miller, C. S. ; Smith, A. J. ; Pettit, F. K. ; Bowie, J. U. ; Eisenberg, D. : The Database of Interacting Proteins: 2004 update. In: *Nucleic acids research* 32 (2004), S. D449–D451

[149] Santo, M. D. ; Foggia, P. ; Sansone, C. ; Vento, M. : A large database of graphs and its use for benchmarking graph isomorphism algorithms. In: *Pattern Recognition Letters* 24 (2003), Nr. 8, S. 1067–1079

[150] Sauer, U. : Metabolic networks in motion: 13C-based flux analysis. In: *Molecular systems biology* 2 (2006), S. 62

[151] Schmith, J. ; Lemke, N. ; Mombach, J. C. ; Benelli, P. ; Barcellos, C. K. ; Bedin, G. B.: Damage, connectivity and essentiality in protein-protein interaction networks. In: *Physica A: Statistical Mechanics and its Applications* 349 (2005), Nr. 3-4, S. 675 – 684

[152] Schreiber, F. ; Schwöbbermeyer, H. : Frequency Concepts and Pattern Detection for the Analysis of Motifs in Networks. In: *Transactions on Computational Systems Biology* 3 (LNBI 3737) (2005), S. 89–104

[153] Schreiber, F. ; Schwöbbermeyer, H. : MAVisto: a tool for the exploration of network motifs. In: *Bioinformatics* 21 (2005), Nr. 17, S. 3572–3574

[154] Schuetz, R. ; Kuepfer, L. ; Sauer, U. : Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. In: *Molecular Systems Biology* 3 (2007), S. 119

[155] Sedgewick, R. : *Algorithms in C++, Parts 1-4: Fundamentals, Data Structure, Sorting, Searching.* Addison-Wesley Professional, 1998

[156] Sedgewick, R. : *Algorithms in C++ Part 5: Graph Algorithms.* Addison-Wesley Professional, 2001

[157] Shamir, R. ; Tsur, D. : Faster Subtree Isomorphism. In: *J. Algorithms* 33 (1999), Nr. 2, S. 267–280

[158] Shen-Orr, S. S. ; Milo, R. ; Mangan, S. ; Alon, U. : Network motifs in the transcriptional regulation network of *Escherichia coli*. In: *Nature Genetics* 31 (2002), Nr. 1, S. 64–68

[159] Slater, P. J.: Maximin Facility Location. In: *Journal of National Bureau of Standards* 79B (1975), S. 107–115

[160] Sporns, O. ; Kötter, R. : Motifs in brain networks. In: *PLoS Biology* 2 (2004), Nr. 11, S. e369

[161] Stephenson, K. A. ; Zelen, M. : Rethinking centrality: Methods and examples. In: *Social Networks* 11 (1989), S. 1–37

[162] Tittmann, P. : *Graphentheorie.* Fachbuchhaus Leipzig im Carl Hanser Verlag, 2003

[163] Ullmann, J. R.: An Algorithm for Subgraph Isomorphism. In: *Journal of the ACM* 23 (1976), Nr. 1, S. 31–42

[164] UniProt Consortium, T. : The Universal Protein Resource (UniProt). In: *Nucleic Acids Research* 35 (2007), S. 193–197

[165] Valente, T. W. ; Foreman, R. K.: Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network. In: *Social Networks* 20 (1998), S. 89–105

[166] Vitkup, D. ; Kharchenko, P. ; Wagner, A. : Influence of metabolic network structure and function on enzyme evolution. In: *Genome Biology* 7 (2006), Nr. 5, S. R39

[167] Wagner, A. ; Fell, D. A.: The small world inside large metabolic networks. In: *Proceedings of the Royal Society London B* 268 (2001), S. 1803–1810

[168] Wasserman, S. ; Faust, K. : *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994

[169] Weise, S. ; Grosse, I. ; Klukas, C. ; Koschützki, D. ; Scholz, U. ; Schreiber, F. ; Junker, B. H.: Meta-All: a system for managing metabolic pathway information. In: *BMC Bioinformatics* 7 (2006), Nr. 465

[170] Wernicke, S. ; Rasche, F. : FANMOD: a tool for fast network motif detection. In: *Bioinformatics* 22 (2006), Nr. 9, S. 1152–1153

[171] Wuchty, S. ; Oltvai, Z. N. ; Barabási, A. L.: Evolutionary conservation of motif constituents in the yeast protein interaction network. In: *Nature Genetetics* 35 (2003), Nr. 2, S. 176–179

[172] Wuchty, S. : Interaction and domain networks of yeast. In: *Proteomics* 2 (2002), Nr. 12, S. 1715–1723

[173] Wuchty, S. : Evolution and topology in the yeast protein interaction network. In: *Genome Research* 14 (2004), Nr. 7, S. 1310–1314

[174] Wuchty, S. ; Stadler, P. F.: Centers of complex networks. In: *Journal of Theoretical Biology* 223 (2003), S. 45–53

[175] Wunderlich, Z. ; Mirny, L. A.: Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains. In: *Biophysical Journal* 91 (2006), S. 2304–2311

[176] Xia, Y. ; Yu, H. ; Jansen, R. ; Seringhaus, M. ; Baxter, S. ; Greenbaum, D. ; Zhao, H. ; Gerstein, M. : Analyzing cellular biochemistry in terms of molecular networks. In: *Annu Rev Biochem* 73 (2004), S. 1051–1087

[177] Yu, H. ; Greenbaum, D. ; Lu, H. X. ; Zhu, X. ; Gerstein, M. : Genomic analysis of essentiality within protein networks. In: *Trends in Genetics* 20 (2004), Nr. 6, S. 227–231

[178] Yu, H. ; Kim, P. M. ; Sprecher, E. ; Trifonov, V. ; Gerstein, M. : The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. In: *PLoS Computational Biology* 3 (2007), Nr. 4, S. e59

[179] Yu, H. ; Luscombe, N. M. ; Lu, H. X. ; Zhu, X. ; Xia, Y. ; Han, J.-D. J. ; Bertin, N. ; Chung, S. ; Vidal, M. ; Gerstein, M. : Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. In: *Genome Research* 14 (2004), Nr. 6, S. 1107–1118

[180] Zhu, D. ; Qin, Z. S.: Structural comparison of metabolic networks in selected single cell organisms. In: *BMC Bioinformatics* 6 (2005), S. 8

[181] Zhu, X. ; Gerstein, M. ; Snyder, M. : Getting connected: analysis and principles of biological networks. In: *Genes & Development* 21 (2007), Nr. 9, S. 1010–1024

A. Zusätzliche Tabellen

Die im Abschnitt 6.5 verwendete Tabelle mit den Resultaten der Zentralitätsanalysen aus verschiedenen Publikationen ist in diesem Anhang abgedruckt.

Referenz	Jeong *et al.* [79]	Jeong *et al.* [79]	Wagner & Fell [167]	Wagner & Fell [167]	Ma & Zeng [111]
Netzwerkart	gerichtet	gerichtet	ungerichtet	ungerichtet	gerichtet
Zentralität	*In-Degree*	*Out-Degree*	*Degree*	*Closeness*	*Degree*
Organismus	*E. coli*	*E. coli*	*E. coli*	*E. coli*	verschiedene
Top 10	H2O	H2O	glutamate	glutamate	Glycerate-3-phosphate
	ADP	ATP	pyruvate	pyruvate	D-Ribose-5-phosphate
	orthophosphate	NAD+	CoA	CoA	Acetyl-CoA
	ATP	NADH	2-oxoglutarate	glutamine	Pyruvate
	NAD+	ADP	glutamine	acetyl CoA	D-Xylulose 5-phosphate
	NADH	orthophosphate	aspartate	oxoisovalerate	D-Fructose 6-phosphate
	L-glutamate	NADPH	acetyl CoA	aspartate	5-Phospho-D-ribose 1-diphosphate
	CoA	L-glutamate	phosphoribosylPP	2-oxoglutarate	L-Glutamate
	NH4+	NADP+	tetrahydrofolate	phosphoribosylPP	D-Glyceraldehyde 3-phosphate
	NADP+	CoA	succinate	anthranilate	L-Aspartate

Tabelle A.1.: Top-10 Metaboliten basierend auf Resultaten verschiedener Publikationen. Angegeben ist jeweils die Art des verwendeten Netzwerkes (gerichteter vs. ungerichteter Graph), die verwendete Zentralität (siehe Kapitel 2 für die Definitionen), der betrachtete Organismus und die Liste der Top-10 Metaboliten aus dieser Publikation. Alle angegebenen Namen der Metaboliten stammen aus der entsprechenden Publikation, eine Vereinheitlichung oder Übersetzung erfolgte nicht. (Fortsetzung nächste Seite.)

Referenz	Ma & Zeng [110]	Ma & Zeng [110]
Netzwerkart	gerichtet	gerichtet
Zentralität	*Out-Closeness*	*In-Closeness*
Organismus	*E. coli*	*E. coli*
Top 10	Pyruvate	Pyruvate
	2-Dehydro-3-deoxy-6-phospho-D-gluconate	Acetyl-CoA
	Acetyl-CoA	Malate
	Glyceraldehyde 3-phosphate	Acetate
	Serine	Formate
	Acetaldehyde	Fumarate
	2-Deoxy-D-ribose 5-phosphate	2-Dehydro-3-deoxy-6-phospho-D-gluconate
	Cystine	Citrate
	Malate	Acetaldehyde
	Phosphoenolpyruvate	Methylglyoxal

Tabelle A.1.: Top-10 Metaboliten basierend auf Resultaten verschiedener Publikationen. Angegeben ist jeweils die Art des verwendeten Netzwerkes (gerichteter vs. ungerichteter Graph), die verwendete Zentralität (siehe Kapitel 2 für die Definitionen), der betrachtete Organismus und die Liste der Top-10 Metaboliten aus dieser Publikation. Alle angegebenen Namen der Metaboliten stammen aus der entsprechenden Publikation, eine Vereinheitlichung oder Übersetzung erfolgte nicht. (Fortsetzung nächste Seite.)

Referenz	Ma & Zeng [110]	Arita [14]	Diese Dissertation
Netzwerkart	gerichtet	gerichtet	gerichtet
Zentralität	*Overall-Closeness*	*Degree*	$\sum$ Max-Fluss-Zent
Organismus	*E. coli*	*E. coli*	*E. coli*
Top 10	Pyruvate	carbon dioxide	Glycerinaldehyd-3-phosphat
	Acetyl-CoA	pyruvate	D-Fruktose 6-phosphat
	Malate	acetyl CoA	1,3-Bisphosphoglycerat
	2-Dehydro-3-deoxy-6-phospho-D-gluconate	ATP	Acetyl CoA
	Acetate	D-glucose	3-Phospho D-Glycerat
	Acetaldehyde	L-glutamate	Pyruvat
	Glyceraldehyde 3-phosphate	D-galactose	L-Malat
	Phosphoenolpyruvate	CoA	Fumarat
	D-4-Hydroxy-2-oxoglutarate	S-adenosyl L-methionine	Malonyl CoA
	Methylglyoxal	D-5-phosphoribosyl-1P	Coenzym A

Tabelle A.1.: Top-10 Metaboliten basierend auf Resultaten verschiedener Publikationen. Angegeben ist jeweils die Art des verwendeten Netzwerkes (gerichteter vs. ungerichteter Graph), die verwendete Zentralität (siehe Kapitel 2 für die Definitionen), der betrachtete Organismus und die Liste der Top-10 Metaboliten aus dieser Publikation. Alle angegebenen Namen der Metaboliten stammen aus der entsprechenden Publikation, eine Vereinheitlichung oder Übersetzung erfolgte nicht.

Erklärung

Selbständigkeitserklärung (entsp. Promotionsordnung §5, Absatz 2b): Hiermit erkläre ich, dass ich diese eingereichte Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

.....
Dirk Koschützki