Eiden, Michael

Vom Fachbereich VI Geographie / Geowissenschaften
der Universität Trier zur Verleihung des
akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte

Dissertation

Integrative Analyse experimenteller Metabolitzeitreihen unter Verwendung von theoretischen Netzwerktopologien und Transkriptomdaten. Eine Studie im systembiologischen Kontext von *Corynebacterium glutamicum*.

Betreuer und erster Berichterstatter:	Prof. Dr. Wolfhard Symader
Zweiter Berichterstatter:	Prof. Dr. Brunhilde Blömeke
Datum der mündlichen Aussprache:	15. Mai 2009
Erscheinungsort und -jahr:	Trier, 2010

Zusammenfassung

Der Forschungsbereich der Systembiologie hat sich in den letzten Jahren mit unvergleichlicher Dynamik entwickelt und sich als interdisziplinäres Feld in den Biowissenschaften etabliert. Die Systembiologie verfolgt hierbei unter anderem das Ziel, biologische Systeme als Ganzes zu betrachten. Die analytische Erfassung der Stoffwechselzwischenprodukte, auch Metaboliten genannt, eröffnet hierbei neue Möglichkeiten. Metaboliten - als chemische Verbindungen vergleichbar niedrigen Molekulargewichtes definiert - stellen Zwischenprodukte *in vivo* ablaufender biochemischer Reaktionen dar. Die in biologischen Systemen - sei es auf der Ebene von einzelnen Zellen, Geweben oder Organismen - ablaufenden Reaktionen werden ihrerseits durch spezialisierte Proteine, die Enzyme, katalysiert. Die enzymatische Aktivität wiederum wird maßgeblich durch transkriptionelle und (post-) translationale Prozesse reguliert, steht folglich also in Abhängigkeit zur im Genom verankerten Information. Hieraus wird die Bedeutung der Metaboliten in der systembiologischen Betrachtungsweise deutlich: sie stellen gewissermaßen Endprodukte jener Vorgänge dar, welche auf der Ebene des Transkriptom und Proteoms gesteuert und ermöglicht werden. Aus dieser Abhängigkeit heraus wird deutlich, dass das Metabolom als Gesamtheit der Metaboliten vergleichsweise näher mit dem Phänotyp des betrachteten Systems verbunden ist, als beispielsweise das Transkriptom oder Proteom. In dieser Arbeit wurden Zeitreihen von Metabolitkonzentrationen untersucht, welche im Rahmen von Fermentationsexperimenten mit dem nicht-pathogenen Bodenbakterium *Corynebacterium glutamicum* erfasst worden sind. Die Fermentationsexperimente wurden auf unterschiedlichen Ausgangssubstraten durchgeführt, wobei die Metaboliten in äquidistanten zeitlichen Abständen erfasst wurden. Zur Korrektur von Messfehlern und zur optimalen Vorverarbeitung der Daten wurde ein maßgeschneidertes System der Datenprozessierung entwickelt. Die unüberwachte Datenstrukturanalyse basierend auf den Metabolitzeitreihen ergab, dass sich die Metaboliten ihrer zeitlichen Ausprägung nicht uniform oder gar zufällig verhalten, sondern sich in Gruppen unterschiedlichen Prozessverhaltens einordnen lassen. Diese unüberwachte Eingruppierung anhand der in den Zeitreihen vorhandenen Strukturen erlaubte eine erste grundlegende funktionelle Zuordnung der Metaboliten. Übergeordnet betrachtet, konnten in den Konzentrationsdaten Strukturen gefunden werden, welche deutliche Übereinstimmungen mit den physiologischen Phasen des bakteriellen Wachstums zeigten und zur Feststellung führten, dass sich der gesamte Stoffwechsel von *C. glutamicum* während der Fermentationsexperimente grundlegend verändert. Die Analyse der Metabolomdaten wurde in einem

nächsten Schritt durch eine theoretische Betrachtungsweise erweitert. Hierzu wurde der Stoffwechsel von *C. glutamicum* rechnergestützt modelliert. Zu diesem Zweck wurde eine Genomannotation durchgeführt, mit dem Ziel einen möglichst umfangreichen und qualitativ hochwertigen Katalog über das enzymatische Repertoire von *C. glutamicum* aus Sequenzinformation abzuleiten. Zusätzlich zur sequenzbasierten Suche nach Enzymen wurden weiterführende organismenspezifische Informationen aus spezialisierten Datenbanken extrahiert. Wissen über vorhandene Enzyme wurde in biochemische Reaktionen übersetzt, welche zu Reaktionsnetzwerken zusammengefügt wurden. Die erzeugten Reaktionsnetzwerke wurden unter Verwendung graphentheoretischer Ansätze analysiert, wobei Netzwerktopologien in Form von Deskriptoren abgeleitet wurden. Die integrative Analyse von experimentellen und theoretischen Deskriptoren ergab, dass sich Eigenschaften von Metabolitzeitreihen deutlich topologischen Merkmalen zuordnen lassen. So zeigt sich beispielsweise, dass ein auffälliger Zusammenhang zwischen der experimentell erfassten Sensitivität im Konzentrationsverlauf eines Metaboliten zu seinem theoretischen Verknüpfungsgrad existiert. Weiterhin konnte gezeigt werden, dass eine hochsignifikante Prozessähnlichkeit zwischen Metaboliten sowohl in direkter Nachbarschaft als auch in größeren Reaktionsabständen auftreten kann, jedoch vorzugsweise dann existiert, wenn beide Metaboliten ihrerseits wenige Reaktionspartner haben. Die integrative Datenanalyse wurde in einem weiteren Schritt abermals erweitert, indem Transkriptominformationen externer Studien integriert wurde. Im Detail wurde in dieser Analyse die Prozessähnlichkeit theoretisch benachbarter Metaboliten des Zentralstoffwechsels in Zusammenschau mit der Transkription enzymkodierender Gene analysiert. Die Ergebnisse zeigten deutlich, dass eine erhöhte Prozessähnlichkeit benachbarter Metaboliten dann existiert, wenn die entsprechenden enzymkodierenden Gene in Abhängigkeit des verwendeten Ausgangssubstrates signifikant exprimiert waren. Nach bisherigem Wissensstand konnte damit erstmals ein Zusammenhang zwischen der Prozessähnlichkeit benachbarter Metaboliten in Abhängigkeit zur Genexpression als Resultat substratinduzierter Anpassungsvorgänge gezeigt werden. Somit konnte im systembiologischen Kontext belegt werden, dass auf der Ebene des Transkriptoms stattfindende Vorgänge sich deutlich bis in die Zeitreiheneigenschaften erfasster Metabolitkonzentrationen durchpausen können. Darüber hinaus konnte gezeigt werden, dass die Berechnung paarweiser Prozessähnlichkeiten das Potenzial zur Charakterisierung der zugrundeliegenden Systemeigenschaften besitzt. So ermöglichte die Betrachtung von Prozessähnlichkeiten aus allen betrachteten Fermentationsexperimenten, signifikante substrat-induzierte Veränderungen als auch invariante Merkmale im Stoffwechsel von *C. glutamicum* zu detektieren.

Abstract

Systems biology has emerged as a tremendously dynamic and interdisciplinary field within biological sciences. It aims at understanding biological systems as a whole instead of investigating well-defined compartments within. The discipline dealing with the identification of metabolites present within biological systems (on the scale of individual cells, tissues or whole multi-cellular organisms), also known as metabolomics now opens the possibility to gain a deeper insight on a system-level. Metabolites are defined as low-molecular weight compounds, representing the intermediates of chemical reactions actually taking place within the system observed. The chemical reactions are catalyzed by enzymes, which represent the most specialized form of proteins. Enzymatic activity however, is a result of transcriptional and (post-) translational processes. This clarifies the importance of metabolites within the system-wide investigation of biological systems: metabolites - in a certain sense - represent end-products of gene regulation and protein activity and therefore are closer to the phenotype of the system observed. This thesis investigated metabolite concentration time-series acquired during fermentation experiments using the non-pathogen organism *Corynebacterium glutamicum*. Fermentation experiments were carried out on different substrates and metabolites were measured in equidistant intervals, resulting in individual time-series of metabolite concentration. A tailored data pre-processing scheme was developed to curate for measurement errors and to enhance the information content of the time-series under investigation. In-depth unsupervised statistical analysis revealed, that metabolites are not behaving uniformly or randomly across time, but instead can be clearly clustered into groups of different temporal behaviour. This finding - solely derived from structures inherent in experimental data - facilitated a first explanation for the position and functional role of metabolites within their metabolic network. Moreover it could be demonstrated that global metabolism is clearly subjected to temporal variations, almost exactly reflecting the physiological phases of bacterial growth, which can also be detected by other means like optical density measurements. The analysis of metabolite time-series properties was extended utilizing a theoretical representation of the organism. Therefore the metabolism of *C. glutamicum* was reconstructed *in silicio*. The annotation of the organisms' genome utilizing up-to-date versions of sequence databases served as a starting point for the computer-based reconstruction. The goal of this approach was to derive a comprehensive and qualitative catalogue on enzymes present in *C. glutamicum*. Additional organism-specific information on enzymes was derived from specialized data-bases and

evaluated with expert knowledge. Subsequently, enzyme information was translated into biochemical reactions, which were merged to reaction networks. The reaction networks created were analyzed using graph-theory based approaches and topology-related descriptors were inferred from the data set. A software system was developed to facilitate the integrative analysis of both experimental and theoretical data. Results revealed, that metabolite time-series properties could clearly be linked to network topologies. For example it could be demonstrated, that the time-series sensitivity is contingent upon the connectivity of the observed metabolite within the theoretical network. Furthermore it showed, that highly significant correlation between metabolite time-series (also called process-similarity in the context of this work) emerges in immediate vicinity as well as across large reaction distances within the network, but - remarkably - is constrained to low mutual connectivity. The integrative data analysis was extended in a second step by incorporating information on transcriptional activity, derived from previous studies, which investigated the same organism under identical experimental conditions. In this context, the process-similarities of neighbouring metabolites within the central metabolism were thoroughly investigated alongside information on transcriptional activity of the corresponding enzyme-coding genes. This analysis was conducted under different substrate conditions, which force the organism to utilize different metabolic pathways for substrate assimilation. Results impressively revealed, that process-similarity between neighbouring metabolites is increased, when the transcription of the corresponding enzyme-coding gene is significantly elevated under the given substrate-induced conditions. Two major findings were inferred from this results. Firstly this is - up to our knowledge - the first time, that mutual metabolite time-series properties could clearly be linked to the underlying transcriptional activity. Moreover it showed, that analysis of pair-wise process-similarities is able to serve as a fingerprint for the underlying system characteristics. In a subsequent step, mutual time-series properties from all fermentation experiments available were analyzed in a combined analysis. This approach clearly unravelled significant substrate-induced alterations as well as conserved features within the metabolism of *C. glutamicum*.

Ich versichere, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Hermeskeil, den 4. Dezember 2008

Für Britta

Danksagung

Die vorliegende Arbeit wurde von Dezember 2005 bis Dezember 2008 in Kooperation mit dem Institut für Biochemie der Universität zu Köln durchgeführt. An dieser Stelle möchte ich jenen Personen meinen Dank aussprechen, ohne die diese Arbeit nicht möglich gewesen wäre.

- Herrn Prof. Wolfhard Symader dafür, dass er mein Interesse an der Arbeit mit komplexen Daten geweckt hat, sowie seine kontinuierliche Betreuung.
- Herrn Prof. Dietmar Schomburg für die freundliche Aufnahme in seiner Kölner Arbeitsgruppe, sowie für die Betreuung vor Ort.
- Frau Prof. Brunhilde Blömeke für die Übernahme des Korreferates.
- Frau Dr. Silke Schrader für die hilfreiche Unterstützung und besonders für das Korrekturlesen der Arbeit.
- Frau Eliane Frimmersdorf für die experimentellen Rohdaten, sowie Herrn Dr. Syed Asad Rahman und Dr. Kai Hartmann für wertvolle Diskussionen.
- Meinen Eltern Josef und Rosie Eiden, sowie meiner Lebensgefährtin Britta Weichmann für ihre Unterstützung und Geduld.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Der Forschungszweig der Systembiologie ist eine vergleichsweise junge Disziplin unter den biologischen Wissenschaften und hat in den letzten Jahren eine unvergleichlich dynamische Entwicklung durchlebt. Ihr Ziel ist es, biologische Systeme tatsächlich als ganzheitliche Systeme zu verstehen und nicht als Agglomeration einzelner biochemischer Komponenten zu betrachten (Kitano, 2002). Zum Verständnis auf Systemebene ist es unter anderem erforderlich, eine Fülle von Daten aus unterschiedlichsten Betrachtungsansätzen zu akquirieren und in einem Gesamtzusammenhang zu betrachten.

Jene analytische Technologie, die hierbei neben der Untersuchung der transkribierten Geninformation (Englisch: „Transcriptomics“) sowie der translatierten Proteine (Englisch: „Proteomics“) zunehmend mehr Verwendung findet, ist die systematische Untersuchung der Zwischenprodukte des Stoffwechsels, der so genannten Metaboliten (Weckwerth, 2003; Kell, 2004). Die Betrachtung der Gesamtheit der Metaboliten (im Englischen auch „Metabolomics“ genannt) ist nicht zuletzt daher in den Fokus der wissenschaftlichen Betrachtung gerückt, da Metaboliten Zwischenprodukte *in vivo* ablaufender biochemischer Reaktionen sind und somit eine Aussage darüber erlauben, welche biochemischen Vorgänge innerhalb des beobachteten Systems (sei es auf der Ebene von Organismen, Geweben oder einzelnen Zellen) zu einem bestimmten Zeitpunkt tatsächlich aktiv sind.

Die Prozesse, die auf der Ebene des Metaboloms stattfinden, sind Resultate jener Vorgänge, die auf der Ebene des Transkriptoms und des Proteoms gesteuert, ermöglicht und durch enzymatische Reaktionen reguliert werden (Sauer et al., 2007). Man könnte sagen, dass die Metabolomforschung deshalb so interessant ist, da sie im Vergleich zur Untersuchung des Transkriptoms oder Proteoms gewissermaßen „näher“ am Phänotyp ist. Hieraus resultiert auch, dass ist in jüngster Zeit die gemeinsame, parallele Betrachtung von Metabolomdaten mit Informationen aus den Bereichen der Transkriptom- oder Proteomforschung wichtiger zur Beantwortung systembiologischer Fragestellungen geworden ist (Fiehn, 2001;

Urbanczyk-Wochniak et al., 2003; Ishii et al., 2007). Dieser Konzeption folgt auch die vorliegende Arbeit, indem sie Informationen aus unterschiedlichen Bereichen zur Beantwortung von Fragestellungen heranzieht und in integrativer Form analysiert (siehe hierzu im Detail Kapitel 2).

Zur analytischen Erfassung des Metaboloms eignen sich zahlreiche Verfahren (Kapitel 4.2) aber besonders massenspektroskopische Ansätze wie zum Beispiel die massenspektroskopisch gekoppelte Gaschromatographie (GC/MS) beziehungsweise die Flüssigchromatographie mit Massenspektrometrie-Kopplung (LC/MS). Jene sind in der Lage, eine robuste, sensitive und über mehrere Größenordnungen hinweg zuverlässige Detektion von Metaboliten zu ermöglichen. Andere analytische Verfahren zur Untersuchung von Metaboliten sind beispielsweise die NMR-Technologie oder vibratonsspektroskopische Verfahren wie die FT/IR- oder Ramanspektroskopie. Anzumerken bleibt allerdings, dass aufgrund der Mannigfaltigkeit der Metaboliten keine analytische Technologie in der Lage ist, sämtliche Metaboliten parallel zu detektieren (Dunn et al., 2005). Daher sollte die Wahl der analytischen Plattform immer unter Berücksichtigung der zu beantwortenden Fragestellung getroffen werden.

Auf bakteriellen Modellorganismen werden weltweit zahlreiche unterschiedliche Versuche zur Untersuchung der Stoffwechselvorgänge durchgeführt (Koek et al., 2006). Die in dieser Arbeit betrachteten experimentellen Daten wurden mit Hilfe der GC/MS Technologie am Modellorganismus *Corynebacterium glutamicum*, im Rahmen des Forschungsbereiches „Metabolomics“ in der Arbeitsgruppe von Prof. Dietmar Schomburg am Institut für Biochemie der Universität zu Köln erhoben. Ziel der Untersuchungen war es, neue Informationen über den Stoffwechsel von *C. glutamicum*, welches in der biotechnologischen Herstellung von Aminosäuren von hoher wirtschaftlicher Bedeutung ist, zu generieren. Im Speziellen wurden hierbei Untersuchungen des Metaboloms zu äquidistanten Zeitpunkten innerhalb von Wachstumsreihen von *C. glutamicum* durchgeführt. Resultat dieser Messungen, welche bei Wachstum des Bakteriums auf verschiedenen Ausgangssubstraten durchgeführt wurden, sind letztendlich Konzentrationszeitreihen individueller Metaboliten. Dieses ist insofern von großer Bedeutung, da zahlreiche Studien, welche den Metabolismus mikrobieller Organismen untersuchten, sich aufgrund des hohen messtechnischen Aufwandes nur auf einen Messzeitpunkt beschränkten und somit eine Prozessbetrachtung bisher nicht ermöglichten. Jene gewonnenen Metabolitzeitreihen und deren Eigenschaften sind die primäre Datengrundlage für die

mathematisch-statistische Analyse dieser Arbeit.

Neben der Untersuchung der experimentellen Daten, wurde eine rechnergestützte Modellierung des Stoffwechsels von *C. glutamicum* durchgeführt. Die Ergebnisse der theoretischen Untersuchung wurden in einem nächsten Schritt in Form einer integrativen Analyse in Zusammenschau mit den experimentellen Ergebnissen untersucht. Darüber hinaus wurden in einem abschließenden Schritt Transkriptomdaten in die Analyse integriert. Die Analyseprozedur besteht folglich aus mehreren aufeinander aufbauenden, komplexer werdenden Schritten. Die vorliegende Arbeit ist nach dem aktuellem Wissenstand die erste Arbeit, welche die Prozessähnlichkeiten von Metaboliten anhand ihrer Konzentrationsverläufe in Zusammenschau mit Netzwerktopologien und Transkriptomdaten untersucht. Zusammenfassend bedeutet dies, dass diese Arbeit mit Hilfe klassischer mathematisch-statistischer Ansätze aus experimentellen metabolischen Daten, theoretischen Netzwerktopologien und Transkriptomdaten neuartiges Wissen generiert, welches dem systemischen Verständnis von komplexen Vorgängen dient.

2 Zielsetzung

Das in dieser Arbeit betrachtete Bakterium *Corynebacterium glutamicum* ist bekannt für seine Anpassungsfähigkeit, die sich unter anderem darin manifestiert, dass es in der Lage ist, selbst auf unterschiedlichsten Nährmedien zu wachsen (Wendisch et al., 2000).

Die messtechnisch erfassbaren Metabolitkonzentrationen aus den Fermentationsexperimenten sind gewissermaßen Resultate jener ablaufenden komplexen regulatorischen Prozesse und erlauben einen Einblick in die systemische Gestalt des Netzwerkes. In dieser Arbeit wurde die Frage geklärt, ob sich in den experimentellen Daten Strukturen und Auffälligkeiten finden lassen, welche Ausdruck der regulatorischen Prozesse sind (Kapitel 7.1.4). Zur Klärung dieser Fragestellung fanden Verfahren der multivariaten Datenstrukturanalyse Anwendung. Da Metabolomdaten Besonderheiten aufweisen, welche sich unter anderem in extrem verschiedenen Konzentrationsverhältnissen äußern (van den Berg et al., 2006), bestand ein generelles Ziel darin, vor der systematischen Datenstrukturanalyse eine geeignete Prozedur der Datenvorverarbeitung (Kapitel 5.1.1) zu entwickeln mit dem Ziel, möglichst viel an Information aus weiterführenden Analysen abzuleiten.

Die darauf aufbauende Fragestellung bestand darin, zu untersuchen, ob sich die in den experimentellen Zeitreihen gefundenen Strukturen mit Hilfe von Zusatzinformationen hinsichtlich ihrer Ausprägungen erklären lassen. Verknüpft man mehrere Informationsebenen miteinander und sucht zwischen diesen nach Zusammenhängen, spricht man, wie in diesem Fall, von einer integrativen Analyse. Die Zusatzinformation wurden in diesem Schritt gänzlich aus einer theoretischen Repräsentation des Organismus abgeleitet, welche organismenspezifisch vorhandenes Wissen über Stoffwechselwege in kondensierter Form repräsentiert. Der Ausgangspunkt hierzu liegt in einer Annotation der vorhandenen Erbinformation (Kapitel 4.3.1). Diese verfolgt das Ziel möglichst aktuelle und verlässliche Erkenntnisse über in *C. glutamicum* vorhandene Enzyme zu erhalten, welche für die Katalyse biochemischer Reaktionen notwendig sind. Aus den gesammelten Erkenntnissen

über vorhandene Enzyme und damit auch über dadurch katalysierende Reaktionen wurden Reaktionsnetzwerke erstellt (Kapitel 4.3.2). Abgeleitete Reaktionsnetzwerke repräsentieren - vereinfacht ausgedrückt - den Stoffwechsel des betrachteten Organismus in virtueller Form. In ihnen sind die Stoffwechselwege und damit die Umsetzungsmöglichkeiten zwischen Metaboliten repräsentiert. Um aus dieser Gesamtheit gültige und biochemisch plausible Stoffwechselwege auch unter der Berücksichtigung der Problematik von Seitenmetaboliten zu erhalten, wurde eine Netzwerkanalyse mit Hilfe graphentheoretischer Ansätze (Kapitel 4.3.3) durchgeführt. Da Metaboliten unterschiedliche Aufgaben und Positionen innerhalb von metabolischen Netzwerken besitzen, bestand ein weitergehendes Ziel dieser Netzwerkanalyse darin, umfangreiche beschreibende Informationen (Deskriptoren) über die Metaboliten und ihre topologischen Eigenschaften innerhalb der untersuchten Netzwerke abzuleiten.

Hierzu gehören - wie im weiteren Verlauf der Arbeit detailliert beschrieben - beispielsweise die Metabolitverknüpfung. Sie beschreibt, in welchem Maße ein Metabolit mit anderen Reaktionspartnern im betrachteten Netzwerk interagieren kann und stellt folglich ein topologische Größe zur Charakterisierung des Nachbarschaftverhältnisses dar. Andere Größen greifen beispielsweise ab, ob ein gegebenes Paar von Metaboliten nach aktuellem Wissenstand enzymatisch katalysiert ineinander umgesetzt werden kann oder nicht. Ist eine Umsetzung möglich, beschreibt der Reaktionsabstand (im weiteren Verlauf der Arbeit auch „Pfadlänge“ genannt) wie viele individuelle Reaktionsschritte hierzu erforderlich sind. Die Gesamtheit der aus den Reaktionsnetzwerken abgeleiteten Informationen wird in den nachfolgenden Kapiteln „Netzwerktopologien“ oder „Topologiedeskriptoren“ genannt.

Die experimentelle und die theoretische Betrachtung von *C. glutamicum* sind voneinander unabhängig und aus unterschiedlichen Gesichtspunkten her motiviert. Sie haben jedoch gemeinsam, dass sie das gleiche Untersuchungsobjekt betrachten.

Um eine integrative Analyse zu ermöglichen, wurde ein neuartiges System zur Datenauswertung von zeitlich aufgelösten Metabolomdaten entwickelt, welches experimentelle Informationen mit theoretischen Netzwerktopologien in Form einer vereinheitlichten Datenstruktur zusammenführt und zwischen den beiden Datensätzen mit Hilfe von Verfahren der multivariaten Statistik nach vorhandenen Zusammenhängen sowie Auffälligkeiten sucht. Die Verknüpfung von zeitlich aufgelösten Metabolitkonzentrationen mit organismenspezifischen theoretischen Netz-

werktopologien wurde in dieser Form noch in keiner wissenschaftlichen Arbeit behandelt.

Zu den Fragestellungen, die im Rahmen der integrativen Analyse untersucht wurden, gehörte beispielsweise auch die Klärung, ob ein Zusammenhang zwischen der Prozessähnlichkeit zweier Metaboliten und ihrem zugehörigen Reaktionsabstand existiert (Kapitel 7.4.2.1). Da hohe Prozessähnlichkeit zwischen Metaboliten ein Hinweis auf eine enge, eventuell noch nicht annotierte Regulation sein kann, wurde dieses Phänomen tiefergehend untersucht.

Als erweiterte Fragestellung - welche gewissermaßen auf den vorherigen Untersuchungen aufbaut - wurde ferner geklärt, ob durch zusätzliche Integration fermentationsspezifischer Transkriptomdaten in das bestehende System, weitergehende Erkenntnisse gewonnen werden können (Kapitel 7.4.3). Hierzu finden Daten aus anderen wissenschaftlichen Studien Anwendung, welche die Genexpression ausgewählter enzymkodierender Gene im Zentralstoffwechsel von *C. glutamicum* unter verschiedenen Fütterungsbedingungen untersuchten (Hayashi et al., 2002; Muffler et al., 2002 oder Gerstmeir et al., 2003). Diese Informationen wurden herangezogen, um Strukturen im zeitlichen Prozessverhalten theoretisch benachbarter Metaboliten tiefergehend zu untersuchen und ihre Ausprägung zu erklären. Zusammengefasst bedeutet dies, dass bei dieser Arbeit Informationen aus drei verschiedenen Ansätzen in Zusammenschau analysiert wurden.

Diese integrative, in der Literatur auch multiparallel genannte Betrachtungsweise ist von hohem Interesse, da sie häufig für ein tiefergehendes, systemisches Verständis des zugrundeliegenden Systems unerlässlich ist (Fiehn, 2001; Kell, 2004).

3 Stand der Forschung

Die Anwendungsgebiete der Metabolomforschung und die dabei verwendeten Methoden sind enorm vielfältig, so dass an dieser Stelle nur die wichtigsten, grundlegenden Ansätze und Ideen gegenübergestellt werden sollen. Einen umfassenden Überblick, vor allem über die aktuellen analytischen Ansätze, die dabei verwendeten Technologien, sowie deren Limitationen und Potenziale, liefert die umfangreiche Studie von Dunn et al. aus dem Jahre 2005.

3.1 Experimentgestützte Untersuchung metabolischer Systeme

Ein Aspekt, der häufig Beweggrund für die experimentelle Untersuchung metabolischer Systeme ist, ist die funktionelle Genanalyse (Bino et al., 2004). Hierbei werden bestimmte Gene gezielt in so genannten „Knock-Out"-Experimenten ausgeschaltet. Dadurch provozierte Veränderungen im Metabolom lassen sich durch den Vergleich des Wildtyps mit der Mutante detektieren. Die gefundenen Unterschiede lassen unter Umständen Rückschlüsse darüber zu, welche Funktion das Gen im Stoffwechsel wahrnimmt und helfen folglich, dass Wissen über regulatorische Mechanismen zu erweitern. Dieser Vorgehensweise folgend, haben beispielsweise schon vor einigen Jahren Tweeddale et al. (1998) Pionierarbeit geleistet, indem sie unter Minimalbedingungen „Knock-Out"-Experimente bei *Escherichia coli* durchführten. Dabei konnten sie feststellen, dass durch das Ausschalten des Gens *RpoS* (welches bei *Escherichia coli* bei Nahrungslimitierung eine Rolle spielt) Veränderungen in der Zusammensetzung des Metaboloms zwischen der RpoS-Mutante und dem Wildtyp existierten, die - zumindest teilweise - mit bereits bekannten Funktionen des mutierten Gens in Beziehung gesetzt werden konnten. Die Verwendung von „Knock-Out"-Experimenten hat schnell in den Bereich der industriellen Biotechnologie Einzug gehalten. Ein Hintergedanke ist hierbei, durch Eingriffe auf

genetischer Ebene möglichst ertragreiche Mutanten, beispielsweise von biotechnologisch verwertbaren Mikroorganismen zu erhalten. Hier ist neben dem in dieser Arbeit untersuchten Bakterium *Corynebacterium glutamicum* exemplarisch die Bierhefe *Saccharomyces cerevisiae* zu nennen. Neben der Untersuchung mikrobieller Metabolome für biotechnologische Zwecke, findet die Metabolomanalyse ferner rege Anwendung im Bereich der experimentellen Pflanzenphysiologie. Ziele sind hierbei unter anderem Ertragssteigerungen bei Nutzpflanzen oder Grundlagenforschung an transgenen Pflanzen. Zu den wichtigsten - auch wegweisenden Arbeiten - gehören hierbei die Untersuchungen an *Arabidopsis thailana* von Fiehn et al. (2000), an *Solanum tuberosum* von Roessner et al. (2000), an oder etwa an *Perilla frutescens* von Yamazaki et al. (2003).

Die Ansätze der experimentellen Metabolomforschung gehen allerdings auch über den bloßen Vergleich von Mutante gegen Wildtyp hinaus. Die Untersuchungen von Steuer et al. (2003) und Weckwerth et al. (2004) waren beispielsweise die ersten, die auf systematische Art und Weise das wechselseitige Verhalten von Metaboliten auf Basis experimentell erhobener Konzentrationsdaten untersucht und beschrieben haben. Hierbei wurden erstmals Korrelationen zwischen Metabolitkonzentrationen als Ähnlichkeitsmaß tiefergehend untersucht und die Zusammenhänge in Form von so genannten Korrelationsnetzwerken graphisch dargestellt. Weiterführende Untersuchungen hierzu wurden beispielsweise von Camacho et al. (2005) geliefert, indem sie in ihrer Studie die Ursache auffälliger Korrelationen zwischen Metaboliten untersuchten. Die genannten Arbeiten verwendeten hierzu als Ausgangsdaten punktuelle Metabolitkonzentrationen aus mehrfachen Replikaten oder Proben, und nicht - wie im Falle dieser Arbeit - Zeitreihen der Metabolitkonzentration. Darin liegt ein entscheidender Vorteil: durch die Betrachtung der Zeitreihen und ihrer Eigenschaften ist eine prozessorientierte Untersuchung möglich, während die Betrachtung von punktuellen Konzentrationsmessungen aus Replikaten gewissermaßen die mittlere Abhängigkeit der Metaboliten aus der gegebenen Grundgesamtheit beleuchtet.

Diese vorliegende Arbeit findet ihre prinzipiellen Wurzeln in diesem Ansatz des paarweisen Vergleiches, geht jedoch sowohl methodisch als auch konzeptionell darüber hinaus. Bisher ist dem Autor noch keine Studie im Bereich der experimentellen Metabolomuntersuchung bekannt, welche sich intensiv mit der Analyse von Zeitreiheneigenschaften beschäftigt, und diese in Zusammenschau mit Zusatzinformation aus der rechnergestützten theoretischen Betrachtungsweise sowie

unter Verwendung von Transkriptomdaten analysiert. Zur Beschreibung der paarweisen Prozessähnlichkeiten von Metaboliten werden in dieser Arbeit neben der Korrelation mehrere unterschiedliche Deskriptoren verwendet. Diese beschreiben beispielsweise die gegenseitige Formähnlichkeit der Zeitreihen, oder ihre wechselseitigen Trendeigenschaften (Kapitel 5.1.2).

3.2 Mathematische Repräsentation metabolischer Systeme

Eine wachsende Anzahl von Forschergruppen im Bereich Systembiologie beschäftigt sich mit der rechnergestützten Nachbildung und Simulation metabolischer Netzwerke. Die hierbei verwendeten Ansätze sowie die benutzten Methoden sind auch hier vielgestaltig und stetiger Weiterentwicklung unterworfen. Grundsätzlich existieren zwei verschiedene Ansätze, wie metabolische Netzwerke mathematisch repräsentiert werden können.

Der erste Ansatz beschäftigt sich mit der Nachbildung sämtlicher vorhandener Reaktionen und Metaboliten in Form einer so genannten stoichiometrischen Matrix (Schilling und Palsson, 1998). Unter der Grundvoraussetzung, dass sich das betrachtete metabolische System in einem Gleichgewicht befindet, lassen sich in einem „Flux-Balance Analysis“ genannten Ansatz die bevorzugten Flussraten und -richtungen abschätzen (Schilling et al., 2001; Covert et al., 2001; Edwards et al., 2002). Der Vorteil dieses Ansatzes liegt darin, dass ermittelte Ergebnisse anhand von Zusatzinformationen, wie beispielsweise thermodynamischer Randbedingungen weitergehend verfeinert werden können. Der entscheidende Nachteil besteht hingegen darin, dass der durch Parametrisierung und Finden des Lösungsraumes bedingte rechnerische Aufwand sehr groß ist. Die Modellierung des gesamten Stoffwechsels ist bei diesem Ansatz oft nicht möglich, weshalb sich Forschergruppen auf die Modellierung einzelner Komponenten des Stoffwechsels, wie beispielsweise den Zitratzyklus oder die Glykolyse beschränken.

Eine gänzlich andere Möglichkeit, metabolische Netzwerke zu modellieren, besteht darin, dass man graphenbasierte Ansätze (Jeong et al., 2000) verwendet. Hierbei werden beispielsweise die Metaboliten als Knotenpunkte und die Reaktionen als die dazwischenliegenden Verbindungslinien repräsentiert. Die Richtungsabhängigkeit von biochemischen Reaktionen kann hierbei in Form von Rich-

tungsindikatoren berücksichtigt werden. Ist dies der Fall, spricht man von einem gerichteten Graphen. Graphenbasierte Modelle des Stoffwechsels lassen sich weitergehend auf topologische Eigenschaften untersuchen (Ma und Zeng, 2003a). Als Werkzeug zur Untersuchung von Reaktionsnetzwerken existieren inzwischen zahlreiche Programme. Eines ist beispielsweise das im Rahmen dieser Arbeit genutzte und am Cologne University Bioinformatics Center (CUBIC) von Rahman et al. (2005) entwickelte Pathway Hunter Tool (PHT).

Der entscheidende Vorteil des graphenbasierten Ansatzes ist, dass er eine direkte Visualisierung von komplexen metabolischen Zusammenhängen erlaubt und somit eine Interpretation einfacher macht. Ein weiterer entscheidender Vorteil liegt darin, dass die Repräsentation großer, genom-weiter metabolischer Systeme möglich ist und Netzwerktopologien (wie in dieser Arbeit systematisch durchgeführt) gut abzuleiten sind. Des Weiteren können Informationen über neue, bisher nicht identifizierte Enzyme beziehungsweise Reaktionen vergleichsweise schnell in das bestehende Netzwerk eingearbeitet und sich dadurch ergebenden Konsequenzen für das restliche Netzwerk bestimmt werden. Dennoch birgt der graphentheoretische Ansatz auch Nachteile, die gesondert kompensiert werden müssen. Insbesondere der Problematik von Seitenmetaboliten wurde in dieser Arbeit viel Aufmerksamkeit gewidmet (vergleiche hierzu Kapitel 4.3.3).

3.3 Datenbanken und externe Informationsquellen

Ohne die Hinzuziehung externer Informationen sind Arbeiten im systembiologischen Kontext nicht sinnvoll durchzuführen. Im Rahmen dieser Arbeit wurden unterschiedlichste externe Datenquellen verwendet. So fanden beispielsweise im Zuge der rechnergestützten Rekonstruktion Sequenzdatenbanken wie Swiss-Prot, TrEMBL (Bairoch und Apweiler, 2000) sowie ProSite (Hulo et al., 2006) für die Genomannotation (vergl. Kapitel 4.3.1.1) Anwendung. Ferner lieferte die für die Betrachtung metabolischer Systeme wichtige Enzymdatenbank BRENDA (Braunschweig Enzyme Database), (Schomburg et al., 2002), wichtige organismenspezifische Informationen und wurde ferner zur Qualitätsüberprüfung verwendet. Sehr häufig wurde die umfangreiche Datenbank KEGG (Kyoto Encyclopedia of Genes and Enzymes) (Kanehisa et al., 2004) verwendet. So stammt beispielsweise

die Zuordnung von Enzymen zu ihren korrespondierenden individuellen Reaktionen aus der in KEGG hinterlegten LIGAND-Datenbank (Goto et al., 2002). Auch für die in dieser Arbeit verwendete Namenskonvention der Metaboliten und Enzyme liefert KEGG den Standard. Die im Bereich der Systembiologie sehr frequentierte Datenbank MetaCyc (Krieger et al., 2004; Caspi et al., 2006) lieferte wie in Kapitel 4.3.1.4 beschrieben Informationen über mutmaßliche Lücken im metabolischen Netzwerk von *C. glutamicum*. Generell wurde die Hinzuziehung von Informationen aus externen Datenquellen im Rahmen dieser Arbeit konsequent unter Zuhilfenahme von Expertenwissen überprüft.

Im Kontext des aktuellen Standes der Forschung ist festzuhalten, dass die vorliegende Arbeit sich sowohl mit experimentellen Daten als auch der theoretischen metabolischen Untersuchung von *C. glutamicum* beschäftigt. Dies, ihre breit aufgestellte Datengrundlage und ihre integrative Konzeption der Datenanalyse machen diese Arbeit in ihrer Konzeption bisher einzigartig.

4 Material und Methoden

Der Stoffwechsel des Bakteriums *Corynebacterium glutamicum* wird, wie bereits erwähnt, anhand experimenteller Daten als auch durch rechnergestützte Modellierungen analysiert. Dieses Kapitel beschreibt im Detail das Vorgehen bei der Durchführung beider Ansätze, die dabei verwendeten Werkzeuge und Methoden. Eine graphische Darstellung, die einen Überblick über die Konzeption beider Ansätze bis hin zur gemeinsamen mathematisch-statistischen Analyse beinhaltet, kann in Abbildung 4.1 gefunden werden. In Kapitel 4 werden die Punkte behandelt, die in der schematischen Darstellung mit weiß unterlegten Kästchen dargestellt sind. Dies betrifft sowohl auf experimenteller und theoretischer Seite die Generierung der Datensätze, bis diese in ihrer rohen, unvorverarbeiteten Form vorliegen. Das nachfolgende Kapitel 5 widmet sich der Weiterverarbeitung der erzeugten Datensätze und leitet geeignete Deskriptoren aus ihnen ab (dargestellt durch grau unterlegte Kästchen). Die integrative Datenanalyse - in einer zweiten Ausbaustufe ergänzt durch externe Transkriptomdaten - folgt aufbauend auf diesen Schritt (dargestellt durch hellbraune Kästchen). Die mathematisch-statistischen Verfahren, die in der integrativen Datenanalyse Verwendung finden, sind in Kapitel 6 dargestellt.

Bevor jedoch im Detail auf das weitere Vorgehen eingegangen wird, soll zunächst ein Überblick über *Corynebacterium glutamicum* selbst gegeben werden.

4.1 Beschreibung von *Corynebacterium glutamicum*

Corynebacterium glutamicum (ATCC 13032) ist ein im Boden vorkommendes, gram-positives nicht pathogenes Bakterium. Taxonomisch gesehen gehört es zur Familie der Actinomyceten, wozu unter anderem auch bekannte Mikroorganismen wie die Krankheitserreger *Corynebacterium diphtheriae*, *Mycobacterium le-*

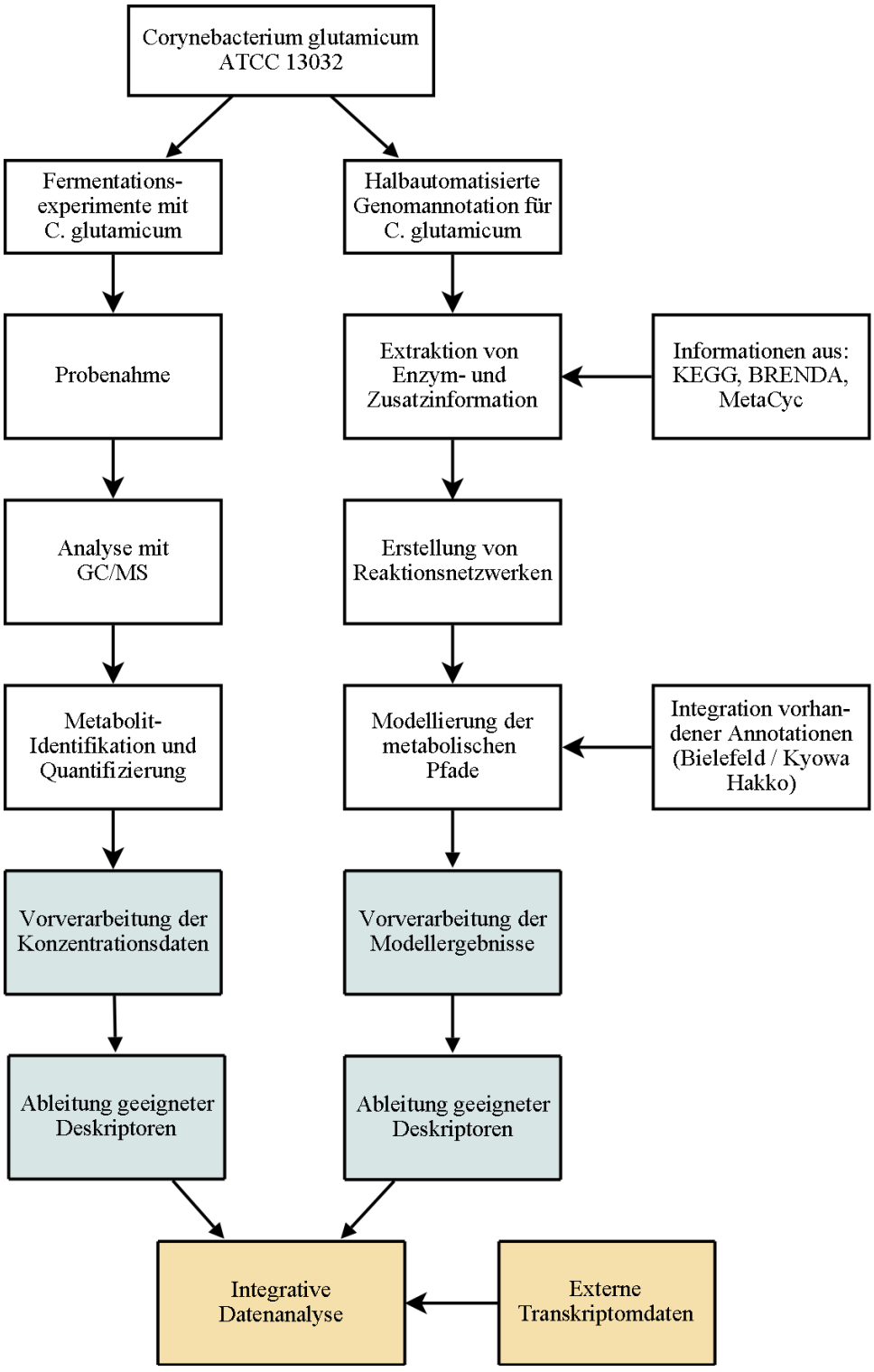

Abbildung 4.1: Schematische Darstellung des Arbeitsablaufes

prae und *Mycobacterium tuberculosis* gehören. Wegen seiner Nicht-Pathogenität, vor allem aber wegen seiner Eigenschaft als universeller Nährstoffverwerter ist *C. glutamicum* besonders leicht zu kultivieren und deshalb auch in der Wissenschaft als Untersuchungsobjekt sehr beliebt (Eggeling und Bott, 2005). Neben der Verwendung als wissenschaftliches Untersuchungsobjekt besitzt *C. glutamicum* auch hohe wirtschaftliche Bedeutung, so wird es unter anderem zur biotechnologischen Herstellung von Lysin oder Glutamat, welches Liebhabern fernöstlicher Küche nur allzu gut als Geschmacksverstärker bekannt sein wird, verwendet. Die Fähigkeit von *C. glutamicum* auf unterschiedlichsten Nährmedien zu wachsen (Liebl, 1991; Wendisch et al., 2000; Gerstmeir et al., 2003) ist ein Indiz für seine Fähigkeit, auf metabolischer Ebene flexibel zu reagieren. Dieser enormen Anpassungsfähigkeit liegt eine ausgesprochene metabolische Robustheit zugrunde, die *C. glutamicum* besonders interessant für eine metabolische Untersuchung macht. Die Untersuchung des Organismus ist in den letzten Jahren nicht unerheblich von dem Biochemischen Institut der Universität zu Köln, dem Institut für Genetik der Universität Bielefeld, sowie dem Institut für Biotechnologie des Forschungszentrums Jülich vorangetrieben worden. Von dort stammt auch die nachfolgende Abbildung 4.2, welche eine rasterelektronenmikroskopische Aufnahme des Bakteriums zeigt. In der Abbildung wird deutlich, dass *C. glutamicum* seinen Namen aufgrund seiner leicht keulenartig (coryneform) verdickten Zellmorphologie erhalten hat.

Die vollständige Entschlüsselung des Genoms von *C. glutamicum* konnte durch Kalinowski et al. im Jahre 2003 abgeschlossen werden. Das Genom, welches nur auf einem einzigen ringförmigen Chromosom (siehe Abbildung 4.3) lokalisiert ist, besitzt rund 3,2 Millionen Basenpaare, wobei 2993 proteinkodierende Gene entdeckt werden konnten. Nur geringfügig später wurde unabhängig von der Bielefelder Arbeitsgruppe vom japanischen Biotechnologiekonzern Kyowa Hakko eine zweite Annotation veröffentlicht (Ikeda und Nakagawa, 2003).

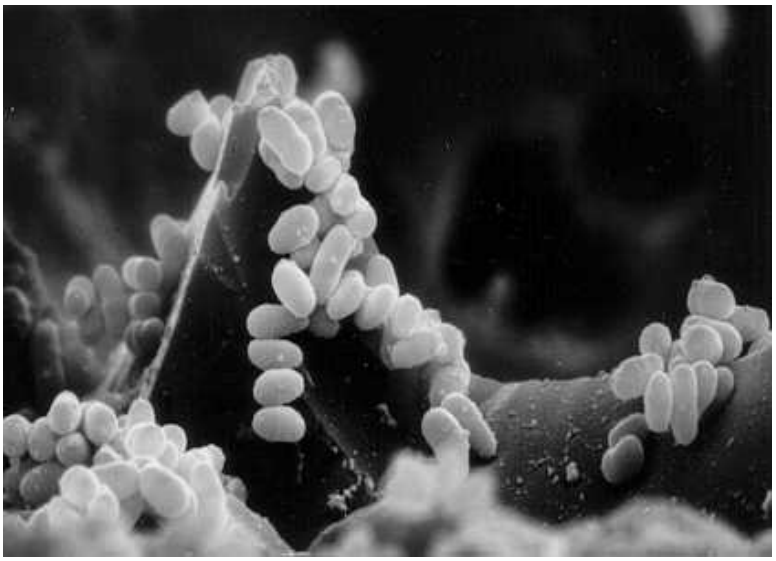

Abbildung 4.2: Rasterelektronenmikroskopische Aufnahme von *C. glutamicum* ATCC 13032 Wildtyp (Quelle: Forschungszentrum Jülich)

4.2 Experimentelle Untersuchung von *C. glutamicum*

Die dieser Arbeit zugrunde liegenden Daten wurden am Institut für Biochemie der Universität zu Köln in der Arbeitsgruppe von Prof. Schomburg erhoben. Die betrachteten Experimente wurden von Frau Eliane Frimmersdorf im Rahmen ihrer Diplomarbeit (Frimmersdorf, 2005) konzipiert und durchgeführt. Als Messverfahren wurde hierbei die kombinierte Gaschromatographie / Massenspektroskopie (GC/MS) angewandt. Diese Technik eignet sich aufgrund ihrer Robustheit und Empfindlichkeit besonders gut zur Untersuchung von Metaboliten, da diese in sehr vielgestaltiger chemischer Struktur und stark unterschiedlichen Konzentrationsverhältnissen auftreten können (Dunn et al., 2005 sowie Goodacre et al., 2004).

Wie erwähnt, wurden mit *C. glutamicum* so genannte Fermentationsexperimente durchgeführt, bei denen eine Zellkultur auf unterschiedlichen Nährmedien (Acetat, Fructose, Glucose, Glutamin, und Lactat) herangezüchtet worden ist. Unter allen Ausgangssubstraten ist *C. glutamicum* in der Lage, zu wachsen, wenngleich sich die Wachstumsraten stark unterscheiden. In der von Strelkov et al. (2004) entwickelten experimentellen Vorgehensweise konnten rund 1000 Verbindungen detektiert werden, wovon 330 signifikant nachzuweisen waren. Von diesen 330

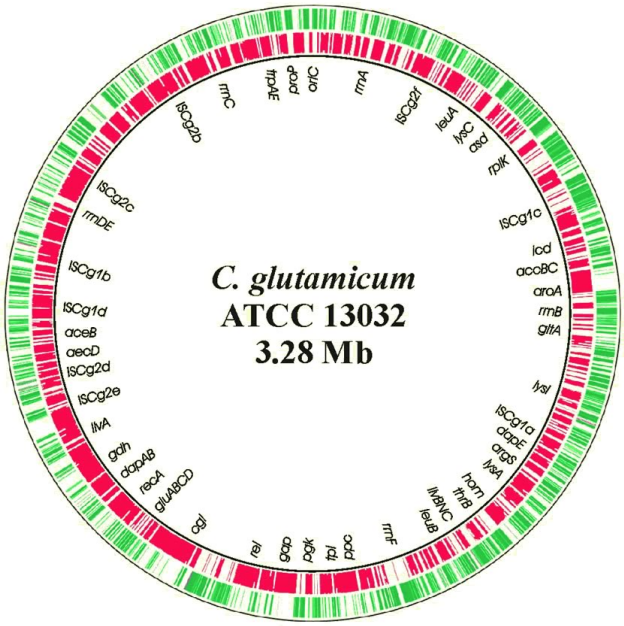

Abbildung 4.3: Chromosomenkarte von *C. glutamicum* ATCC 13032 Wildtyp. Die äußeren und inneren Ringe zeigen die Open Reading Frames an, welche im Uhrzeigersinn, respektive gegenläufig dazu, abzulesen sind. Die Position ausgewählter Gene ist gekennzeichnet. (Quelle: Institut für Genetik der Universität Bielefeld)

konnten wiederum 164 identifiziert und 121 unterschiedlichen Metaboliten zugeordnet werden.

4.2.1 Probenahme

Zellkulturen von *Corynebacterium glutamicum* (ATCC 13032 Wildtyp) wurden auf den bereits erwähnten Nährmedien aerob herangezüchtet. Die Probenahme fand während der Durchführung der Fermentationsexperimente zu exakt definierten Zeitpunkten im Abstand von 60 Minuten statt. Hintergrund dieses Vorgehens war es, eine Aussage über den Metabolismus von *C. glutamicum* zu unterschiedlichen Zeitpunkten, das bedeutet innerhalb unterschiedlicher Wachstumsphasen (Lag-Phase, exponentielle und stationäre Phase), zu erhalten und somit eine dynamische Betrachtung zu ermöglichen. Vom biologischen Standpunkt aus betrachtet, kann davon ausgegangen werden, dass der Organismus in der Lag-Phase auf die veränderten Umweltbedingungen durch Anpassung reagiert. Danach hat der Organismus seinen Stoffwechsel soweit angepasst, dass ein exponentielles Wachstum und damit die vordringliche Produktion von Biomasse möglich ist. Nach einer kurzen Übergangsphase folgt die stationäre Phase, in der das Hauptsubstrat aufgebraucht ist und kein weiteres Zellwachstum mehr stattfindet. In der anschließenden Absterbephase überwiegt der Abbau bereits produzierter Biomasse. Parallel zur Probenahme wurde das Zellwachstum anhand der der optischen Dichte (OD) bei einer Wellenlänge von 600 nm erfasst. Die nachfolgende Abbildung 4.4 zeigt den Verlauf der optischen Dichte während eines Fermentationsexperimentes von *C. glutamicum* mit Glucose als Ausgangssubstrat. Die jeweiligen physiologischen Wachstumsphasen können hieraus - wie folgt - bestimmt werden. Die Lag- sowie die Übergangsphase erstrecken sich vom Beginn des Fermentationsexperimentes bis ca. 420 Minuten. Es ist davon auszugehen, dass in dieser Phase vornehmlich Anpassungsvorgänge stattfinden. Die exponentielle Phase, welche sich durch sehr starke Biomasseproduktion und rasche Zellteilung kennzeichnet, kann ungefähr in dem Bereich von 480 bis 720 Minuten nach Beginn des Experimentes angesiedelt werden. Nach einer kurzen Übergangsphase, die sich von 720 bis 780 Minuten erstreckt, folgt die stationäre Phase, in der die optische Dichte ein Plateau erreicht.

Bei der Probenahme wurden für jeden Zeitpunkt drei Proben an Zellextrakt mit Hilfe eines Entnahmerohrs entnommen. Das Volumen der Proben wurde der jeweiligen optischen Dichte angepasst, so dass $5x10^{10}$ Zellen entnommen wurden. Die

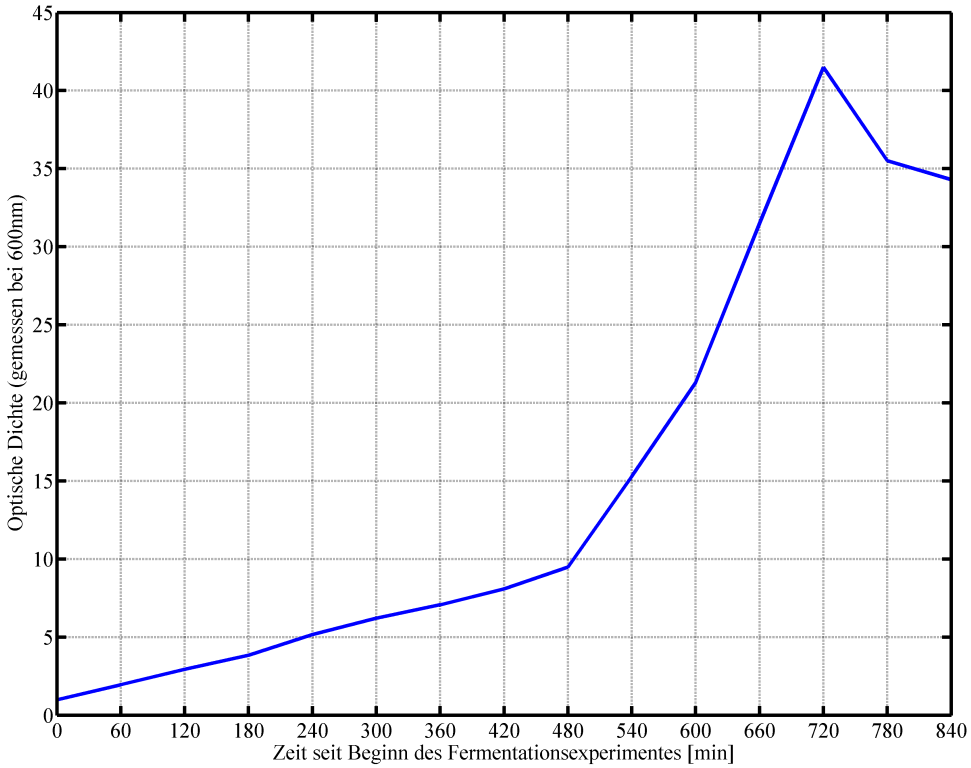

Abbildung 4.4: Verlauf der optischen Dichte bei 600 nm, gemessen bei einem Fermentationsexperiment von *C. glutamicum* auf Glucose

gewonnene Zellsuspension wurde anschließend vom Medium durch Zentrifugation befreit und nachfolgend einer Waschprozedur unterzogen. Die Waschung wurde zweimalig durchgeführt, die Waschlösung anschließend ihrerseits durch Zentrifugation abgetrennt, sodass von der Probe ein sogenanntes Zellpellet für die weitergehenden Arbeitsschritte verblieb.

Um die Metaboliten aus der Probe zu extrahieren, wurde das Zellpellet in 1,5 ml Methanol + 40 μl/ml Ribitollösung resuspendiert und in einem Ultraschallbad behandelt. Das verwendete Methanol und die mechanische Einwirkung durch den Ultraschall dienen hierbei der Zerstörung der Zellwände (Strelkov et al., 2004). Das im gleichen Arbeitsschritt hinzugegebene Ribitol spielt jedoch eine andere Rolle: es wird als interner Standard hinzugegeben, welcher in einem späteren Schritt - der GC/MS-Analyse - wegen seines charakteristischen Auftretens im Chromatogramm - eine wichtige Rolle bei der Normierung der Daten spielt (Strelkov et al., 2004). Zusätzlich wurde die Probe anschließend mit destilliertem Wasser intensiv durchmischt, danach mit Chloroform versetzt und erneut durchmischt. Zur Phasentrennung der Metaboliten wurde das Gemisch bei 6000 Umdrehungen pro Minute und Raumtemperatur für 6 Minuten zentrifugiert. Die hydrophile (bzw. polare) Phase des Gemischs wurde anschließend entnommen und durch Erwärmung unter einem Abzug von noch enthaltendem Methanol befreit. Hiernach wurden die Proben abschließend getrocknet und bei -20° Grad Celsius gelagert. In der vorliegenden Arbeit wurde ausschließlich die polare Phase für die anschließende GC/MS-Analyse verwendet. Eine Erfassung der Metaboliten ist folglich nie gänzlich unbeeinflusst von der Wahl des experimentellen Vorgehens.

4.2.2 GC/MS-Technologie

Keine momentan zur Verfügung stehende analytische Methode erlaubt die wünschenswerte simultane Detektion aller auftretender Metaboliten (Dunn et al., 2005) in einer Probe. Dies hängt in erster Linie von der Vielgestaltigkeit der auftretenden Metaboliten, das heißt deren Masse, Polarität, Volatilität sowie der jeweiligen gerätetechnischen Eigenschaften ab. Im Rahmen dieser Arbeit wurden sämtliche Untersuchungen mit Hilfe der Gaschromatographie-Massenspektroskopie durchgeführt, da diese für metabolische Fragestellungen hinsichtlich ihrer Sensitivität, ihrer großen Messdynamik sowie ihrer Anwendbarkeit für eine Vielzahl von Metaboliten besonders geeignet ist. Die Verwendung der GC/MS-Technologie in der

Metabolituntersuchung geht in Bezug auf pflanzenphysiologische Aspekte zurück bis in die frühen 1970er Jahre, wo sie systematisch von Horning und Horning (1971) angewendet wurden. Weitere entscheidende Impulse wurden u.a. von Sauter et al. (1991) in den frühen 1990er Jahren geliefert.

Bei dieser Technologie werden im Prinzip die beiden Technologien Gaschromatographie und Massenspektroskopie miteinander gekoppelt. Diese Kopplung der Technologien hat für die Detektion von Metaboliten den Vorteil, dass sie eine Identifikation gewissermaßen zweifach erlaubt. Sie erfolgt über die charakteristische Retentionszeit des Metaboliten (das heißt der Zeit bis zum Auftreten des Signals im Chromatogramm) und zum anderen über das dazugehörige charakteristische Massenspektrum. In der chromatographischen Säule findet die Auftrennung der Metaboliten beziehungsweise deren Derivate statt. Diese Auftrennung erfolgt in einer zeitlichen Abfolge und ist in erster Linie abhängig von der Größe, Struktur und Flüchtigkeit der betrachteten Substanzen, sowie experimentellen Parametern wie insbesondere der Säulenlänge und -temperatur sowie der Flussrate des verwendeten Trägergases. Die zu untersuchenden Proben werden nach dem von Dr. Sergey Strelkov (2004) im Rahmen seiner Doktorarbeit entwickelten Methode „coryPTV“ vermessen. Hierbei erfolgte die Injektion des Probenmaterials in den Gaschromatographen programmgesteuert und temperaturabhängig mit Hilfe eines so genannten PTV- (programmed temperature vapourizer) Injektors. Im Anschluss an die chromatographische Auftrennung erfolgte die Analyse im integrierten Massenspektrometer. Hierbei werden die Verbindungen infolge eines extrem starken elektromagnetischen Feldes beschleunigt, in ihre fragmentspezifischen Bestandteile zerlegt und detektiert.

4.2.3 Metabolitidentifikation und -quantifizierung

Wie im vorhergehenden Kapitel beschrieben, liefert die GC/MS-Analyse gewissermaßen eine zeitlich aufgelöste Abfolge von Massenspektren, deren Summation den so genannten Totalionenstrom (total ion current, TIC) ergibt. In der zeitlichen Betrachtung ist der Verlauf des Totalionenstromes auch unter der Bezeichnung Chromatogramm bekannt. Nachfolgende Abbildung 4.5 zeigt ein beispielhaftes Chromatogramm für eine, aus einem Experiment mit *C. glutamicum* untersuchten Probe.

Wie bereits erwähnt erlaubt die GC/MS-Analyse die Identifikation von Meta-

boliten auf zweierlei Weise. Rechnerisch wird sie durch einen Abgleich der beobachteten Peaks bei den jeweiligen m/z-Verhältnissen an den dazugehörigen Retentionszeiten mit in Datenbanken abgespeicherten Informationen durchgeführt. Da sich bei der GC/MS-Technologie einzelne Signale zeitlich gesehen häufig zu einem zusammengesetzten Signal überlappen können, müssen gegebenenfalls überlappende Peaks in Ihre Einzelbestandteile aufgeteilt werden. Dieses Verfahren nennt sich Dekonvolution und wird zusammen mit der Identifikation der Metaboliten durch das vom National Institute of Standards und Technology (NIST) in Gaithersburg / USA entwickelte Programmpaket AMDIS (Automated Mass Spectral Deconvolution and Identification System), welches von Stein im Jahre 1999 entwickelt wurde, durchgeführt. AMDIS ist mit Hilfe des Verfahrens der Datendekonvolution in der Lage, aus überlappenden Massenspektren anhand der Elutionsprofile „reine“ Spektren zu berechnen, die für die anschließende Suche in spektralen Datenbanken verwendet werden können. Die Verwendung von AMDIS dient der Identifikation der Metaboliten und ihrer Derivate; die anschließende quantitative Bestimmung der Metaboliten und ihrer Derivate findet unter Anwendung des Softwarepaketes Xcalibur 1.2 (entwickelt vom Gerätehersteller Thermo Finnigan in San Jose / USA) statt.

Die Metabolitquantifizierung wird durch die Integration der nach der Dekonvolution vorliegenden Peakflächen erreicht. Diese Integration muss häufig anhand von Expertenwissen manuell am Bildschirm durchgeführt und kontrolliert werden. Anschließend liegen die Metabolitkonzentrationen in so genannten „Pseudo-Amount“-Werten, also gewissermaßen semi-quantitativ vor. Die Schritte der Identifikation und der Quantifizierung wurden für jede Messung eines Triplikates separat durchgeführt werden. Nach der Normierung durch den internen Standard Ribitol und nach Addition aller zu einem Metaboliten gehörenden Derivate, erhält man die (Pseudo-) Konzentrationen für jeden identifizierten Metaboliten zum betrachteten Zeitpunkt der Probenahme.

Führt man diese GC/MS-Analyse während eines Fermentationsexperimentes mehrfach durch, so kann man durch Aneinanderreihung der jeweiligen Konzentrationswerte einen zeitlichen Ablauf des Konzentrationsverlaufs für den betrachteten Metaboliten erhalten. In der nachfolgenden Abbildung 4.6 ist exemplarisch der Konzentrationsverlauf zweier Metaboliten unter Fütterungsbedingungen von *C. glutamicum* mit Glucose dargestellt. Da allerdings die Massenspektroskopie nicht in der Lage ist, zwischen Stereoisomeren zu unterscheiden, kann es vor-

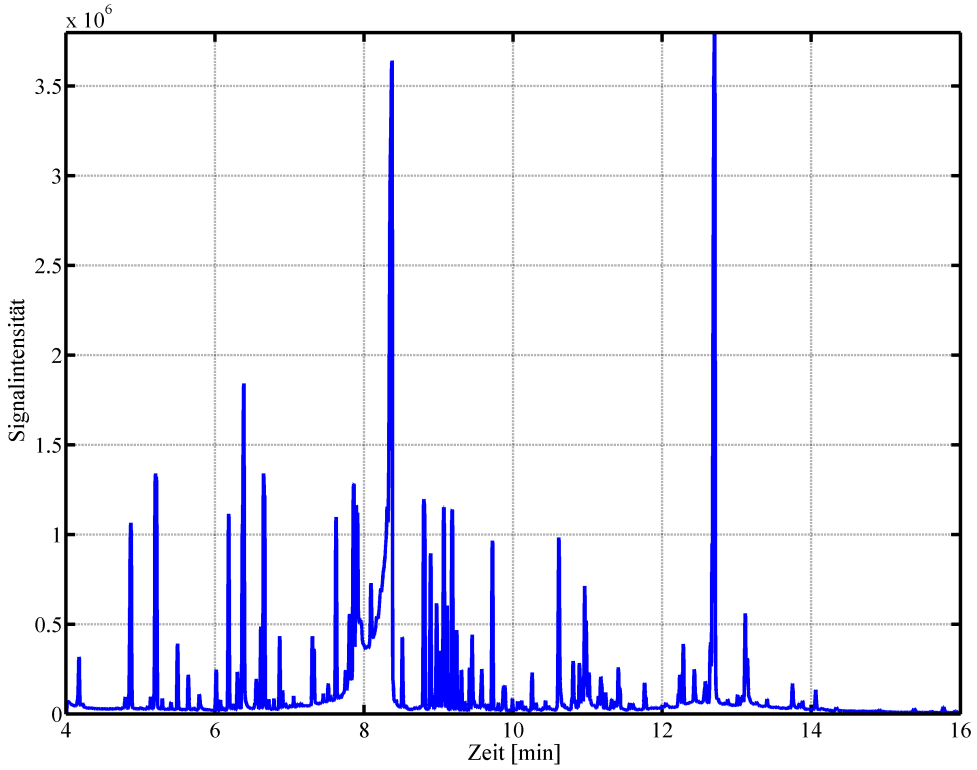

Abbildung 4.5: Beispielhaftes Chromatogramm einer Untersuchung von *C. glutamicum* ATCC 13032 Wildtyp, (Quelle: Institut für Biochemie, Universität zu Köln)

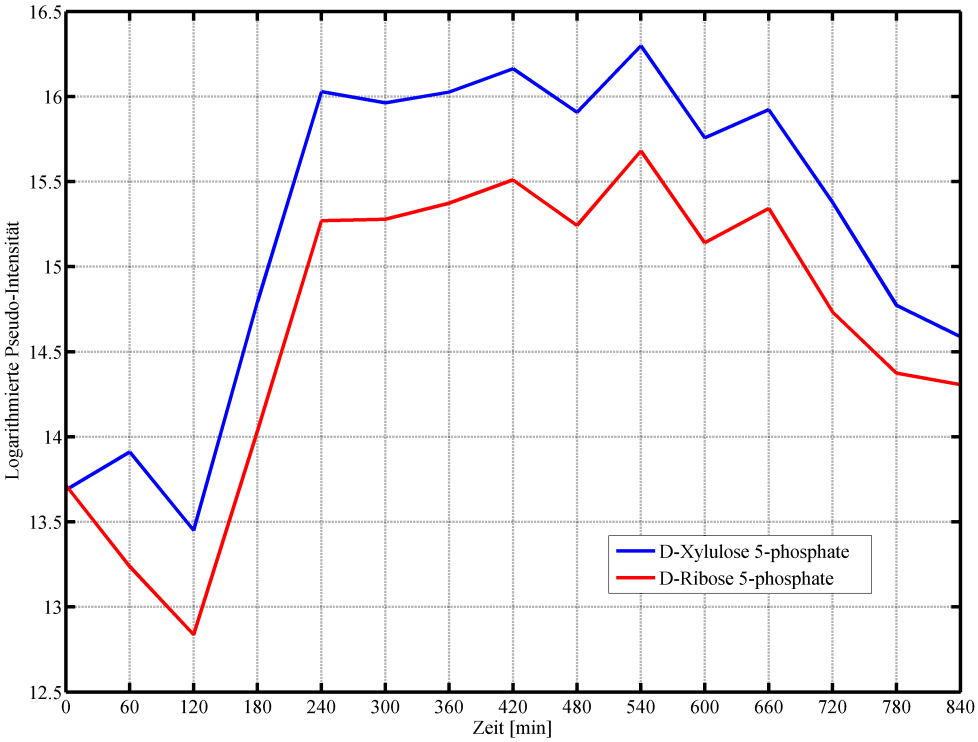

Abbildung 4.6: Exemplarische Konzentrationszeitreihen der Metaboliten D-Xylulose 5-phosphate (C00231) und D-Ribose 5-phosphate (C00117) während der Fermentation von *C. glutamicum* mit Glucose.

kommen, dass eine Zuordnung von Konzentrationen zu einem Metaboliten nicht eindeutig durchgeführt werden kann. Deshalb wurden zunächst die Bezeichnungen aller potenziell auftretenden Stereoisomere beibehalten und in einem späteren Schritt (vergleiche Kapitel 5.1.1.2) anhand von Zusatzinformationen aus der Datenbank KEGG organismenspezifisch für den Stoffwechsel von *Corynebacterium glutamicum* evaluiert. Die primäre Datenerhebung ist bei der experimentellen Betrachtung von *C. glutamicum* mit diesem Schritt beendet - auf die Verarbeitung der Rohdaten und die Informationsextraktion wird im nachfolgenden Kapitel 5 eingegangen.

4.3 Theoretische bioinformatische Untersuchung von *C. glutamicum*

Ziel der bioinformatischen Untersuchung von *C. glutamicum* war es, dessen Stoffwechsel möglichst vollständig und exakt in einem rechnergestützten Netzwerk abzubilden. Hierzu wurden, wie in Kapitel 3.3 bereits angerissen, zahlreiche Informationsquellen hinzugezogen, um eine möglichst vollständige Betrachtungsweise zu erlangen. Da diese externen Informationsquellen mit den Ergebnissen vieler internationaler Forschergruppen gespeist werden, ändert sich deren Umfang als auch Informationsgehalt ständig. Die Vorgehensweise bei der theoretischen Untersuchung von *C. glutamicum* sei an dieser Stelle kurz skizziert, detaillierte Informationen finden sich in den angegebenen Unterkapiteln. Am Beginn der bioinformatischen Untersuchung steht die grundlegendste aller biologisch verfügbaren Informationen, und zwar die Genomsequenz. Basierend auf dieser Genomsequenz wurde eine Genannotation für *C. glutamicum* durchgeführt, mit dem Ziel, möglichst viele neue Informationen für die anschließende Modellierung des Stoffwechsels zu gewinnen. Hierzu wurde eine von Frau Dr. Urte Wendt im Jahre 2003 entwickelte Annotationsprozedur inhaltlich aktualisiert und programmtechnisch sowie konzeptionell weiterentwickelt; das Vorgehen hierzu ist im nachfolgenden Kapitel 4.3.1 beschrieben. Ziel der Annotationsprozedur war in erster Linie die Detektion putativer enzymkodierender Gene. Diese Information über das enzymatische Repertoire von *C. glutamicum* ist enorm wichtig, denn Enzyme fungieren als Biokatalysatoren für chemische Reaktionen. Kennt man die in einem Organismus vorkommenden Enzyme, so ergibt dies einen ersten Hinweis darauf, welche chemi-

schen Reaktionen im Stoffwechsel des betrachteten Organismus ablaufen können. Die Gesamtheit aller ermittelten Reaktionen wurde anschließend zu Reaktionsnetzwerken zusammengeführt. Die erstellten Reaktionsnetzwerke wurde mit dem Pathway Hunter Tool (PHT), einem von Dr. Syed Asad Rahman am Cologne University Bioinformatics Center entwickelten Softwaretool (Rahman et al., 2005) für die Fragestellungen dieser Arbeit untersucht. Nähere Beschreibungen hierzu finden sich detailliert in Kapitel 4.3.3.

4.3.1 Halbautomatisierte Genomannotation

Um zu verstehen, welche enzymatische Reaktionen potenziell im Stoffwechsel des untersuchten Bakteriums *C. glutamicum* möglich sind, ist Information über das Vorhandensein von Genen, sowie deren funktioneller Rolle von ganz entscheidender Bedeutung. Besonderes Augenmerk wurde bei der Genomannotation darauf gelegt, möglichst viel an zusätzlicher, qualitativ hochwertiger Information zu gewinnen, um bei der anschließenden gemeinsamen Betrachtung mit experimentell verfügbaren Daten, eine möglichst große Schnittmenge zu erzeugen. Das bereits am Institut für Biochemie der Universität zu Köln vorhandene System der halbautomatisierten Genomannotation wurde hierzu grundlegend aktualisiert. Hierbei wurden sowohl programmtechnische Bestandteile wie zum Beispiel verwendete Suchalgorithmen auf den neuesten Stand gebracht, ferner benutzte Datenbanken aktualisiert, als auch Modifikationen im Quelltext vorgenommen, die beispielsweise einen verbesserten Export und Weiterverarbeitung der Ergebnisse ermöglichten. Ziel all dieser Aktualisierungs- und Anpassungsschritte war es, möglichst aktuelle und hochwertige Informationen für *C. glutamicum* zu erhalten. Diese Prozedur der Genomannotation gliedert sich in mehrere aufeinander folgende Schritte und ist nachfolgend beschrieben.

4.3.1.1 Durchführung einer Genvorhersage

Hierbei wird das verfügbare Genom von *C. glutamicum* einer so genannten Genvorhersage unterzogen. Da die Gene in der Regel durch so genannte Start- und Stop-Kodons flankiert sind, ist eine maschinelle, rechnergestützte Vorhersage der Genpositionen möglich. Hierzu wurden 3 unterschiedliche Programme, die zur Vorhersage von Genen in prokaryotischen Zellen entwickelt wurden, verwendet. Im Einzelnen handelt es sich um die Verfahren Glimmer (Delcher et al., 1999 und

2007), GenemarkS (Besemer et al., 2001) sowie FgenesB (Softberry Inc., Mount Kisco / USA, `http://www.softberry.com`). Glimmer nutzt künstliche neuronale Netze sowie Hidden Markov Models zur Vorhersage potenzieller Gene. GenemarkS hingegen koppelt Hidden Markov Models mit einem unüberwachten Auswerteschritt. Für die oben genannten Programme waren zum Zeitpunkt dieser Arbeit neuere, verbesserte Versionen verfügbar, so dass diese in die Prozedur integriert wurden. Die verwendeten Algorithmen ermittelten unabhängig voneinander für *C. glutamicum* knapp über 3000 putative Gene, wobei leichte Unterschiede zwischen den gewählten Algorithmen zu verzeichnen waren. Anschließend wurden die Nukleotidsequenzen der jeweiligen Gene in Proteinsequenzen übersetzt, um eine Datenbanksuche zu ermöglichen.

4.3.1.2 Suche nach korrespondierenden Proteinsequenzen

Für die ermittelten Proteinsequenzen aus der Genvorhersage wurde eine Suche in Sequenzdatenbanken durchgeführt, um Informationen über die Funktionen der gefundenen mutmaßlichen Gene zu erhalten. Dazu wurden die neuesten Versionen der Datenbanken SwissProt und TrEMBL heruntergeladen und verwendet. Im Vergleich zu den im Jahre 2003 verfügbaren Informationen, ist die SwissProt-Datenbank im Umfang von ca. 140000 auf rund 240000 Einträge und die TrEMBL-Datenbank von ca. 1 Mio. auf run. 3,3 Mio. Einträge angewachsen (Stand: Oktober 2006). Diese beiden Datenbanken unterscheiden sich dadurch, dass SwissProt ausschließlich experimentell bestätigte Information enthält, d.h. verlässlichere Informationen liefert, während die TrEMBL-Datenbank ungleich umfangreicher ist, jedoch auch Einträge über mutmaßliche, jedoch noch nicht experimentell bestätigte Genfunktionen und enzymatische Reaktionen enthält. Beide Datenbanken haben jedoch gemeinsam, dass sie aufgrund intensiver weltweiter Forschung in den letzten Jahren stark gewachsen sind.

Die Suche nach Sequenzübereinstimmungen in den oben genannten Datenbanken erfolgte mit BLAST (Basic Local Alignment Search Tool), einem der weltweit am häufigsten eingesetzten Werkzeuge zur Analyse biologischer Sequenzdaten, welches von Altschul et al. im Jahre 1990 entwickelt wurde. Die gefundenen Treffer bei der Suche nach Sequenzübereinstimmungen wurden anschließend nach ihrer Qualität sortiert und einem Bewertungsschema unterzogen. Bei diesem Bewertungsschema werden die jeweiligen Treffer nach ihrer Güte, dem so genannten

„Expectancy-Value“ (kurz: E-Value) sortiert und in Güteklassen eingeteilt. Der E-Value ist ein statistisches Maß, welches die Güte der Übereinstimmung zweier Sequenzen beschreibt. Je kleiner der E-Value ist, desto besser ist die Übereinstimmung der vorgegebenen Sequenz mit der korrespondierenden Sequenz in der Datenbank. Für die jeweiligen mutmaßlichen Gene wurden neben dem besten Treffer auch nächstbessere Treffer extrahiert, sofern sie sich entweder in der gleichen Güteklasse befanden wie der beste Hit oder in der nachfolgenden Klasse zu finden waren. Diese im weiteren Verlauf der Arbeit „Mehrfach-Hit Strategie“ genannte Vorgehensweise diente dazu, auszuschließen, dass weitere gute Sequenzübereinstimmungen verlorengehen. Oft kommt es beispielsweise vor, dass der beste Hit für ein putatives Gen mit einem unvollständigen und damit unspezifischeren Enzymeintrag annotiert ist und nachfolgende Treffer mit geringfügig schlechterem E-Value eine Komplettierung des Enzymeintrages erlauben. Weiterhin kann es auch vorkommen, das Multi-Enzymkomplexe erst durch die Berücksichtigung nachfolgender Treffer vervollständigt werden. Da die Ableitung von enzymatisch katalysierter Reaktionen aus einer Genomannotation vor allem im Hinblick auf die Erstellung der Reaktionsnetzwerke besonders wichtig ist, wurde hierbei großes Augenmerk darauf gelegt, möglichst vollständige und daher spezifische Enzymeinträge zu erhalten.

Die Tabelle 4.1 zeigt ein Beispiel dieses Vorgehens. In der Spalte „putatives Gen“ findet sich eine Auswahl von Einträgen, wie sie vom Genvorhersageprogramm GenemarkS ermittelt wurde. In der Spalte „Enzym mit bestem Hit“ wird die EC-Nummer (Enzymbezeichnung) jener Proteinsequenz aus der Datenbank angezeigt, welches die größte Übereinstimmung mit der vorgegebenen Sequenz aus der Genvorhersage hat. Da es natürlich vorkommen kann, dass bestimmte Gene nicht enzymkodierend sind, muss auch dies berücksichtigt werden. In der Spalte „Alle gefundenen Enzyme“ werden alle EC-Nummern aller Treffer in der jeweiligen Datenbank dargestellt. Es zeigt sich, dass die Anzahl von Treffern je nach Gen stark variieren kann, was unter anderem mit der Länge der jeweiligen Sequenz zusammenhängt. Die nachfolgenden 4 Spalten geben an, wie viele der ermittelten Treffer in den einzelnen Güteklassen zu finden waren. Wie bereits erwähnt wurde als Gütekriterium der Sequenzübereinstimmung der E-Value verwendet. Hierbei wurden im Detail 4 Klassen berücksichtigt. In Klasse eins werden E-Values kleiner E^{-80} berücksichtigt, in Klasse zwei E-Values größer E^{-80} aber kleiner E^{-35}, in Klasse 3 E-Values größer E^{-35} aber kleiner E^{-7} und in der letzten Klasse die E-Values grö-

Tabelle 4.1: Exemplarisches Beispiel für die Zuordnung putativer Gene zu gefundenen Enzymeinträgen aus der BLAST- Suche. Die Anzahl gefundener Treffer ist in vordefinierte Güteklassen der Sequenzübereinstimmung aufgeschlüsselt.

Putatives Gen	Eintrag des besten Hits	Alle gefundenen Einträge	Kl. 1	Kl. 2	Kl. 3	Kl. 4
GM1	kein Enzym	kein Enzym	68	119	14	
GM1001	EC 3.1.11.6	EC 3.1.11.6				13
GM1002	EC 3.1.11.6	EC 3.1.11.6		1		84
GM1003	EC 1.17.1.2	kein Enzym				1
		EC 1.17.1.2	17	89	41	
GM1005	kein Enzym	kein Enzym			8	3
GM1007	kein Enzym	kein Enzym		10	63	4
GM1008	kein Enzym	kein Enzym			5	
		EC 2.3.1.-			17	
GM1009	kein Enzym	kein Enzym	7	5	3	18
GM101	EC 1.5.99.8	kein Enzym			2	
		EC 1.2.1.-			6	3
		EC 1.2.1.16			3	2
		EC 1.2.1.22			1	
		EC 1.2.1.24			8	
		EC 1.2.1.27			4	4
		EC 1.2.1.28			1	
		EC 1.2.1.3			36	18
		EC 1.2.1.36			14	
		EC 1.2.1.39			1	
		EC 1.2.1.47			5	
		EC 1.2.1.5			12	
		EC 1.2.1.65			2	
		EC 1.2.1.68			1	
		EC 1.2.1.71			2	
		EC 1.2.1.8			38	2
		EC 1.2.1.9			4	
		EC 1.5.1.12		21	6	1
		EC 1.5.1.6			4	
		EC 1.5.99.8		4		

ßer E^{-7}. Im Beispiel des putativen Gens GM101 findet sich keine Sequenzübereinstimmung in der ersten sondern erst in der zweiten Güteklasse. Hierbei finden sich die Enzymeinträge EC 1.5.1.12 und EC 1.5.99.8, was den Enzymen „1-Pyrroline-5-Carboxylate Dehydrogenase“ und „Proline Dehydrogenase“ entspricht. Der beste Treffer ohne Berücksichtigung weiterer Einträge („Einfach-Hit Strategie“) wäre in diesem Fall der Enzymeintrag EC 1.5.99.8, obwohl der korrespondierende E-Value für den Enzymeintrag EC 1.5.1.12 nur geringfügig schlechter war und sich zusätzlich in der gleichen Güteklasse befand. Wie bereits erwähnt, konnten häufig durch die „Mehrfach-Hit Strategie“ unvollständige Enzymeinträge (bei denen die letzte Nummer in der Nomenklatur durch ein Auslassungszeichen ersetzt war) komplettiert werden. Dieses zahlt sich insofern aus, das nur komplette EC-Einträge zur Erstellung der Reaktionsnetzwerke verwendet werden können.

Die Anzahl der jeweiligen Treffer pro Enzymeintrag wurde hingegen nicht als Kriterium zur Gewichtung verwendet, da diese Information in erster Linie von der vorherrschenden Forschungsrichtung beeinflusst sein kann. Die BLAST-Suche und das anschließende Bewertungsschema wurden insgesamt sechsfach durchgeführt, je einmal für die benutzten Genvorhersageprogramme Glimmer, Genemark und FgenesB und für die untersuchten Datenbanken SwissProt und TrEMBL. Die Durchführung der Annotationsprozedur nahm mehrere Tage Rechenzeit in Anspruch. Nach erfolgter Datenbanksuche und Gewichtung wurden die Ergebnisse fusioniert. Im Detail wurden zuerst die Ergebnisse der SwissProt- und TrEMBL-Suche für die jeweiligen Genvorversageprogramme individuell zusammengeführt. Um die qualitativen Unterschiede zwischen der SwissProt- und der TrEMBL-Datenbank zu berücksichtigen, wurde ein gewichtetes Auswahlverfahren für die Fusionierung angewendet. Hierbei wurde so vorgegangen, dass für solche Open Reading Frames (ORFs), für die die SwissProt-Datenbank kein Sequenzhomolog lieferte, der korrespondierende beste Treffer aus der TrEMBL-Datenbank, sofern er existiert, entnommen wurde. Bei ORFs hingegen, für die in beiden Datenbanken Ergebnisse vorhanden waren, wurden im Allgemeinen die Ergebnisse der SwissProt-Suche bevorzugt. Notwendige Ausnahmen wurden bei der manuellen Kontrolle der Ergebnisse festgestellt. Sie erfolgten ausschließlich dann, wenn der beste SwissProt-Treffer keinen Enzymeintrag lieferte, der korrespondierende TrEMBL-Eintrag einen Enzymeintrag besaß, einen E-Value von Null aufwies und das entsprechende Enzym als experimentell für *C. glutamicum* bestätigt in der Datenbank BRENDA gefunden wurde. Diese besondere Gewichtung wurde bei

rund 190 der insgesamt rund 3000 ORFs durchgeführt. Vereinfacht ausgedrückt könnte man sagen, dass die Annotation auf SwissProt- Einträgen basiert und nur im Bedarfsfall aufgefüllt wurde. Sämtliche hierzu verwendeten Skripte wurden in MATLAB programmiert.

Nach diesem Schritt lagen für die drei benutzten Genvorhersageprogramme vollständige Listen vor, die sowohl aus SwissProt- und TrEMBL-Einträgen bestanden. Diese drei Listen wurden nun ihrerseits zu einer gemeinsamen Liste fusioniert, indem die Start- und Stop-Positionen der gefunden ORFs auf dem Genom - welche sich für die Vorhersageprogramme unterscheiden können - verglichen und Berücksichtigung von vordefinierten Grenzen der Überlappung zusammengefasst wurden. Anschließend wurden die Endergebnisse der BLAST-Suche in eine MySQL-Datenbank überführt.

4.3.1.3 Integration organismenspezifischer Zusatzinformation

Zusätzlich zur BLAST-Suche nach Sequenzübereinstimmungen in den Proteindatenbanken, die gewissermaßen das Herzstück der Genannotation darstellt, wurde die SwissProt-Datenbank auch nach Schlüsselworten durchsucht. Hierbei wurde gezielt nach dem Suchstring „Corynebacterium glutamicum“ in der Spalte für organismenspezifische Einträge gesucht. Bei den gefundenen ermittelten Einträgen handelt es sich um experimentell bestätigte Einträge, welche direkt an *C. glutamicum* verifiziert werden konnten und damit besonders hoch zu gewichten sind. Da die meisten dieser Einträge auch über die reine Sequenzsuche ermittelt werden konnten, wurde ein Abgleich durchgeführt und es wurden einige zusätzlich Einträge in die erstellte Datenbank integriert.

4.3.1.4 Vergleich der ermittelten Enzyme

Die im vorangegangenen Kapitel beschriebene Genomannotation erbrachte für den untersuchten Organismus *C. glutamicum* eine Liste von 591 beteiligten Enzymen für die „Mehrfach-Hit Strategie“ hervor. Anzumerken ist, dass in der Liste basierend auf der Annotation noch einige Enzymeinträge enthalten sind, die eine unvollständige Nomenklatur aufweisen (wobei ein Auslassungszeichen toleriert wurde). Jene Einträge wurden nach einem Vergleich mit weiteren Informationsquellen - wie nachfolgend beschrieben - eliminiert. Zum Vergleich wurde die generierte Liste zusätzlich mit den Einträgen aus der Enzymdatenbank BRENDA

beziehungsweise AMENDA sowie mit den Ergebnissen von MetaCyc, verglichen. Die Datenbank BRENDA enthält hochqualitative und überprüfte organismenbezogene Enzymeinträge, welche aus der Literatur, Sequenzdatenbanken als auch aus anderen biochemischen Untersuchungsmethoden abgeleitet wurden. Betrachtet man die experimentellen Einträge in der BRENDA-Datenbank, so sind hier Einträge zu finden, welche auf SwissProt, TrEMBL sowie anderen Untersuchungsmethoden beruhen. Für den Vergleich der in BRENDA enthaltenen Enzyminformationen zu der aus der Annotation hervorgegangenen Enzymliste wurde sequentiell vorgegangen. Zuerst wurde überprüft, ob alle in BRENDA enthaltenen Enzymeinträge basierend auf SwissProt auch in der sequenzbasierten Suche gefunden werden konnten. Von den 217 individuellen auf SwissProt basierenden Enzymeinträgen für *C. glutamicum* in der BRENDA-Datenbank konnten alle bis auf 3 Einträge durch die BLAST-Suche auf SwissProt bestätigt werden. Bei den 3 Enzymeinträgen handelt es sich um solche, für die bei der BLAST-Suche in der Zwischenzeit Sequenzübereinstimmungen höherer Qualität (selbstverständlich organismenbezogen) gefunden wurden. Da diese auf der BLAST-Suche basierenden Enzymeinträge redundant auch bei anderen ORFs vorkamen, wurden die bestehenden Einträge aus BRENDA importiert. Vergleicht man jedoch die auf TrEMBL basierenden 323 individuellen Einträge aus der BRENDA-Datenbank mit der BLAST-Suche, so wird es ungleich schwieriger eine qualitative Gewichtung vorzunehmen. TrEMBL-Einträge sind nicht experimentell bestätigt und ihrerseits häufig das Resultat durchgeführter Annotationen. Ein auf der TrEMBL-Datenbank basierender Eintrag ist folglich viel geringer zu gewichten. Er wurde nur dann aus der BRENDA-Datenbank übernommen, wenn er auf dem gleichen Sequenzabschnitt (abgreifbar über die Accession-Number), organismenspezifisch für *C. glutamicum* vollständigere Enzymeinträge lieferte als die BLAST-Suche. Dies war für 10 Einträge der „Mehrfach-Hit Strategie“ der Fall. Insgesamt wurden also 13 Einträge aus der BRENDA-Datenbank übernommen.

AMENDA, als zusätzlicher Bestandteil der BRENDA-Datenbank erlaubt es, organismenspezifische Informationen für *C. glutamicum* anhand von textbasierten Suchverfahren in Literaturdatenbanken durchzuführen. Hierbei ist es allerdings vonnöten, die zugehörigen wissenschaftlichen Publikationen manuell darauf zu überprüfen, ob das betrachtete Enzym tatsächlich Relevanz für *C. glutamicum* besitzt oder nicht. Die AMENDA-Suche lieferte 17 Enzymeinträge, von denen 2 noch nicht integriert waren und als relevant erachtet wurden. Abschließend wur-

de ein Vergleich mit der Datenbank MetaCyc durchgeführt. MetaCyc ist deshalb besonders interessant, da in dieser Datenbank organismenspezifische Informationen über mutmaßliche „Pfadlücken“ enthalten sind (Krieger et al.,2004; Caspi et al., 2006). Hierbei werden mit Hilfe von Suchalgorithmen und vergleichender Netzwerkanalysen Enzyme, welche eine Rolle als Lückenfüller zum vollständigen Funktionieren des Stoffwechsels spielen, ermittelt. Diese Analyse lieferte für *C. glutamicum* eine Liste von 268 potenzieller Enzyme, die potenziell eine solche Funktion als besitzen könnten. Für 100 von diesen 268 konnte diese Hypothese durch korrespondierende putative Gensequenzen untermauert werden. Von diesen 100 „Lückenfüllern“ waren 38 bereits durch die BLAST-Suche integriert, sodass 62 zusätzliche Enzyme letztendlich aus der MetaCyc-Betrachtung übernommen werden konnten.

Nach reiflicher Überlegung und um dem Zustand Rechnung zu tragen, dass Informationen aus Quellen wie AMENDA sowie MetaCyc unter Umständen stärker auf Annahmen beruhen, wurde beschlossen, die Erstellung des Reaktionsnetzwerkes (auch vitueller Organismus genannt) für *C. glutamicum* in zwei Varianten vorzunehmen. Die konservative Variante „VGL1“ berücksichtigt im nachfolgenden die 604 individuellen Enzymeinträge die aus der Genomannotation mit anschließender BLAST-Suche sowie der Abfrage der BRENDA-Datenbank hervorgegangen sind. Sämtliche dieser Einträge wurden von Hand kontrolliert. Die erweiterte Variante „VGL2“ berücksichtigt darüber hinaus zusätzlich Informationen aus der AMENDA- sowie der MetaCyc-Datenbank und ist mit 668 individuellen Enzymen deutlich umfangreicher. Die Entscheidung, die Erstellung des Reaktionsnetzwerkes in zwei Varianten durchzuführen, von denen sich erstere eng auf der Genomsequenz und an experimentell bestätigte Informationen hält, während die zweite zusätzliche, hypothetische Informationen integriert, beruht auch auf weiteren Gründen. In der späteren integrativen Datenanalyse kann nur die Schnittmenge der experimentellen und theoretischen Daten untersucht werden. Insbesondere die Auswirkungen bei der Verwendung der Informationen aus MetaCyc, beispielsweise auf die Anzahl gefundener Stoffwechselwege, soll hierbei untersucht werden. Nachfolgende Tabelle 4.2 gibt einen Überblick darüber, wie sich die Enzymkataloge für die beiden Varianten VGL1 und VGL2 zusammensetzen.

Vergleicht man dem Enzymumfang der bestehenden Annotationen von Kalinowski et al. 2003 und Ikeda und Nakagawa 2003 mit den aus der Untersuchung dieser Arbeit hervorgegangenen Varianten VGL1 und VGL2, so fällt auf, dass der

Tabelle 4.2: Übersicht der ermittelten individuellen Enzymeinträge und ihrer Herkunft für die Reaktionsnetzwerke VGL1 und VGL2

	BLAST-SUCHE	zusätzliche Informationen aus:			SUMME
		BRENDA	AMENDA	MetaCyc	
VGL1	591	13	0	0	604
VGL2	591	13	2	62	668

Enzymumfang nicht unerheblich gesteigert werden konnte (siehe Tabelle 4.3). Die Zunahme um rund 50 Enzyme (entspricht etwa 10%) bei der konservativen Variante des virtuellen Organismus ist in erster Linie auf den gestiegenen Informationsgehalt der Sequenzdatenbanken SwissProt und TrEMBL seit der Publikation der vorhandenen Annotationen zurückzuführen in den Jahr 2003 zurückzuführen. Der deutlich höhere Enzymumfang der erweiterten Variante ist - wie bereits erwähnt - auch auf die Integration hypothetischer Zusatzinformationen zurückzuführen. Schwerpunktmäßig wurde im weiteren Verlauf dieser Arbeit auf die konservative Variante VGL1 eingegangen.

Tabelle 4.3: Annotationsspezifische Anzahl individueller Enzyme

Bielefeld (CGB)	Kyowa Hakko (CGL)	(VGL1)	(VGL2)
554	538	604	668

4.3.2 Erstellung der Reaktionsnetzwerke

Die aus der Annotationsprodezur hervorgegangenen Enzymlisten wurden mit Hilfe der LIGAND-Datenbank (Goto et al., 2002) der KEGG-Plattform in Reaktionsnummern übersetzt. Vereinfacht ausgedrückt bedeutet dies, dass Wissen über das Vorhandensein von Enzymen in reaktionsspezifische Informationen übersetzt wurde. Die Reaktionen alleine reichen jedoch nicht zur Erstellung eines metabolischen Netzwerkes aus. Sie stellen - wenn man sich das Netzwerk als graphische Darstellung vorstellt - die Kanten zwischen den Knotenpunkten, die durch die Metaboliten repräsentiert werden, dar. Deshalb war es in einem nächsten Schritt notwenig, diejenigen Metaboliten zu definieren, zwischen denen die gefundenen

Reaktionen ablaufen dürfen. Zu Lösung dieser Fragestellung wurden zwei unterschiedliche Ansätze durchdacht. Erstens die Verwendung strikt organismenspezifischer Informationen über das theoretische Vorhandensein von Metaboliten, wie sie in der KEGG-Datenbank für *C. glutamicum* hinterlegt sind. Dieses Vorgehen hätte jedoch den entscheidenden Nachteil, dass neue Erkenntnisse aus der Genomannotation wieder verloren gehen, da die Metabolit-Information in KEGG auf den alten Annotationen beruhen. Das gesamte Prozedere der Genomannotation mit neuen Suchalgorithmen etc. wäre somit sinnlos. Aus diesem Grunde wurde diese Idee verworden. Die zweite Überlegung bestand darin, den generellen Referenzstoffwechsel, wie er ebenfalls in der KEGG-Datenbank hinterlegt ist, heranzuziehen. Dieser enthält organismenübergreifende Informationen über sämtliche bekannten Reaktionswege und Enzyme und setzt sich aus zahlreichen Untersuchungen verschiedenster Organismen zusammen. Dieser Ansatz hat jedoch ebenfalls einen entscheidenden Nachteil, dass er Metaboliten berücksichtigt, die in höheren Organismen oder Pflanzen vorkommen und deshalb untypisch für Bakterien sind. Aus diesem Grunde wurde ein Mittelweg gegangen, indem die Gemeinsamkeit der in der Gattung *Corynebacterium* vorhandenen Metaboliten als Referenz verwendet wurde. Hierzu gehören Informationen aus verwandten Stämmen wie beispielsweise *C. efficiens* oder *C. jeikeium*. Dieser Schritt hat den entscheidenden Vorteil, das neue Informationen aus der Genomannotation integriert werden können und zeitgleich sichergestellt ist, dass die Metaboliten auch in tatsächlich bei *C. glutamicum* beziehungsweise nah verwandten Organismen der gleichen Gattung vorkommen. Wie die spätere Netzwerkanalyse (Kapitel 7.2.2) zeigen wird, hat sich dieses Vorgehen doppelt ausgezahlt. Es konnte nämlich gezeigt werden, dass in KEGG vorgehaltene Informationen über einzig in *C. glutamicum* vorkommende Metaboliten teilweise unvollständig sind.

Vordefinierte Reaktionsnetzwerke zahlreicher Organismen sind im Pathway Hunter Tool in Form von Textdateien im Programmverzeichnis hinterlegt. Die aus der Genomannotation hervorgegangenen Reaktionsnetzwerke wurden mit Hilfe eines (für diesen Zweck in MATLAB geschriebenen Programms) in das entsprechende Format konvertiert und als Textdateien in das Programmverzeichnis des Pathway Hunter Tools kopiert. Das Pathway Hunter Tool dient unter anderem der Berechnung der kürzesten Stoffwechselwege zwischen zwei Metaboliten innerhalb eines vordefinierten metabolischen Netzwerkes. Nach der oben beschriebenen Bereitstellung der neuen Reaktionsnetzwerke im PHT standen insgesamt vier verschiedene

Reaktionsnetzwerke für *C. glutamicum* zur Verfügung: basierend auf der Bielefelder Annotation (CGB), der Kyowa Hakko-Annotation (CGL) sowie in den beiden Varianten (VGL1 und VGL2) - hervorgegangen aus dieser Arbeit.

Zu Testzwecken wurde eine Schnellanalyse im Pathway Hunter Tool durchgeführt. Diese ermöglicht es, metabolische Netzwerke miteinander zu vergleichen und erste Kenngrößen abzuleiten. Hierbei fällt auf, dass eine deutlich höhere Komplexität in den neuen Netzwerken gefunden werden kann (vergleiche Tabelle 4.4). Die gesteigerte Anzahl der Reaktionen resultiert in erster Linie aus dem erweiterten Wissen über das enzymatische Repertoire des betrachteten Organismus. Die Anzahl der Metaboliten resultiert aus der gestiegenen Anzahl von Reaktionen und der Verwendung des bakteriellen Referenzorganismus.

Tabelle 4.4: Vergleich verwendeter Reaktionsnetzwerke für *C. glutamicum*. Standardparameter für die Schnellanalyse im Pathway Hunter Tool: lokale Molekülähnlichkeit 15%, globale Molekülähnlichkeit 1%, gerichtete Pfade, Mapping Algorithmus: KEGG

Parameter	(CGB)	(CGL)	(VGL1)	(VGL2)
Anzahl Enzyme	554	538	604	668
Anzahl Reaktionen	907	889	1435	1520
Anzahl Metaboliten	1069	1075	1557	1604

Um für die nachfolgenden Analysen ferner einen detaillierten Vergleich zwischen den verschiedenen virtuellen Varianten von *C. glutamicum* zu erhalten, wurden sämtliche Netzwerkmodellierungen vierfach, das heisst sowohl für Reaktionsnetzwerke der Bielefelder Annotation (CGB), der Kyowa Hakko-Annotation (CGL) sowie für beide Varianten des neuen aus der Genannotation hervorgegangenen virtuellen Organismus (VGL1 und VGL2) durchgeführt.

4.3.3 Modellierung der Stoffwechselwege und Ableitung von Netzwerktopologien

Zum organismenspezifischen Finden von metabolischen Pfade innerhalb gegebener Reaktionsnetzwerke benötigt das PHT neben der Definition des Start- und Endmetaboliten (entspricht dem Edukt und dem Produkt der gesamten Reaktionskette) weitere Startparameter. Hierzu gehören: die globale und lokale Molekül-

ähnlichkeit. Sie beschreiben jeweils die molekulare Ähnlichkeit zwischen zwei Metaboliten, wobei sich die globale Ähnlichkeit auf die Ähnlichkeit zwischen Edukt und Produkt der Reaktionskette bezieht und die lokale Ähnlichkeit jene zwischen Edukt und Produkt der Einzelreaktionen betrachtet. Globale und lokale Molekülähnlichkeit können als Schwellenwerte betrachtet werden. Dies bedeutet, dass nur dann metabolische Pfade gefunden werden, wenn die Ähnlichkeitswerte oberhalb des definierten Schwellenwertes liegen. Dieses Vorgehen der Nutzung eines molekularen Ähnlichkeitsmaßes, dient unter anderem dazu, Wege über sehr kleine - und daher sehr unähnliche - Metaboliten in der theoretischen Betrachtungsweise zu unterbinden. Als zusätzliche Programmeinstellung erlaubt es das PHT einzuschränken, ob die Richtung der Reaktion Berücksichtigung findet. Dies bedeutet, dass entweder nur nach gerichteten, oder ungerichteten Reaktionen gesucht wird. Zusätzlich kann der so genannte „Mapping-Algorithmus“, welcher bestimmt, wie Reaktionspartner untereinander verknüpft werden dürfen, ausgewählt werden. Dies ist insofern wichtig, da in in den meisten betrachteten Reaktionen mehr als ein Edukt in mehrere Produkte umgewandelt wird. Von daher gilt es zu klären, wie die entsprechenden Reaktionspartner miteinander in Verbindung stehen. Es stehen mit „KEGG“ und „CUBIC“ zwei Varianten zur Auswahl. Der „KEGG“-Algorithmus orientiert sich - wie der Name schon vermuten lässt - an der KEGG-Datenbank, während das „CUBIC“-Mapping von Dr. Syed Asad Rahman entwickelt wurde, um zusätzliche Verknüpfungswege zwischen Metaboliten zu identifizierten. Das CUBIC-Mapping ist dadurch charakterisiert, dass es im Vergleich zum KEGG-Mapping eine deutlich höhere Anzahl von Verknüpfungen zwischen Metaboliten erlaubt.

Zum Testen des Pathway Hunter Tools und zum Finden der optimalen Einstellungen für die nachfolgende organismenweite Modellierung wurden zahlreiche Testläufe berechnet. Für die Vorversuche wurden folgende Einstellungen gewählt: lokale Ähnlichkeit variierend zwischen 15 und 35% Prozent, globale Ähnlichkeit variierend zwischen 1 und 5 % unter Verwendung gerichteter Pfade. Ebenfalls verändert wurden die Einstellungen bezüglich des Mapping-Verfahrens und des verwendeten Reaktionsnetzwerkes (vergl. Kapitel 4.3.2).

Erste Analysen auf den Vorversuchen zeigten, dass in den ermittelten metabolischen Pfaden trotz Verwendung der lokalen und globalen Ähnlichkeit, nicht beabsichtigte Wege über so genannte Seitenmetaboliten auftraten. Eine feststehende Definition für Seitenmetaboliten existiert allerdings in der noch jungen Wissen-

schaft der Metabolomforschung nicht, obwohl erste Ansätze geschaffen wurden (Ma und Zeng, 2003a). In dieser Arbeit werden als Seitenmetaboliten solche Metaboliten definiert, die entweder in sehr hoher Konzentration in der Zelle auftreten und daher in keinem ursächlichen Zusammenhang zur metabolischen Regulation stehen (in diesem Fall spricht man auch von gepoolten Metaboliten), oder eine extrem hohe Verknüpfungszahl zu anderen Metaboliten aufzeigen. Zusätzlich wurden ferner flüchtige Metaboliten wie beispielsweise CO_2 oder Metaboliten, welche beispielsweise Elektronen oder funktionelle Gruppen transportieren, auf der Basis einzelner Reaktionen individuell berücksichtigt.

Die nachfolgende Tabelle 4.5 stellt die im Rahmen dieser Arbeit definierten Seitenmetaboliten dar. Hierzu wurden die Metaboliten nach ihrer Verknüpfungsanzahl (welche aus einem globalen Referenzorganismus abgleitet wurde) absteigend angeordnet. Es zeigt sich, dass sehr kleine und häufig auftretende Metaboliten eine hohe Verknüpfungzahl aufweisen, das heisst an vielen Reaktionen beteiligt sind. Angemerkt sei an dieser Stelle, dass für die Bezeichnung der Metaboliten in dieser Arbeit aufgrund der fehlenden Vereinheitlichung die englischsprachige Nomenklatur aus der KEGG-Datenbank verwendet wurde. Diese hat den Vorteil, dass sie eine eindeutige Zuordnung der Substanz anhand der Compound-Nummer (C-Nummer) erlaubt.

Da man verhindern möchte, dass das Pathway Hunter Tool fälschlicherweise metabolische Pfade ermittelt, die beispielsweise ihren Weg über flüchtige Metaboliten wie etwa CO_2 oder über ATP nehmen, muss diese Definition gesondert im Programm hinterlegt werden. Hierzu wurde die so genannte „mapped-reaction“-Datei, welche innerhalb des PHT prinzipiell erlaubte Reaktionen definiert, aufwendig manuell angepasst. Nach Abschluss der Arbeiten wurden für die vier betrachteten Reaktionsnetzwerke Modellierungen der Stoffwechselwege unter „CUBIC“- und „KEGG“-Bedingungen durchgeführt. Die metabolische Pfade wurden hierbei zwischen einer vordefinierten Liste von Metaboliten (die gleichermaßen in allen Fermentationen detektiert wurden) analysiert . Diese Modellierungen nahmen jeweils 5-6 Tage Rechenzeit in Anspruch. Das Pathway Hunter Tool erzeugt als Ausgabe bis zu 1,5 Mio. Zeilen lange zusammenhängende Textdateien, in der sämtliche theoretischen Informationen abgelegt sind. Aus diesem Grunde wurde in MATLAB ein Softwaretool entwickelt, welches aus den Ausgabe-Dateien die relevanten Informationen extrahiert (vergleiche hierzu Kapitel 5.2.2). Die primäre Datenerhebung ist bei der theoretischen Betrachtung von *C. glutamicum* mit der

Durchführung der Modellierungen beendet. Auf die Informationsextraktion wird in Kapitel 5.2 eingegangen.

Tabelle 4.5: Übersicht über die als Seitenmetaboliten definierten Metaboliten und deren aus dem generellen Referenzstoffwechsel abgeleiteten Verknüpfungszahlen

Metabolit	C-Nummer	Reaktionsverknüpfungszahl
H_2O	C00001	2120
O_2	C00007	798
H^+	C00080	789
$NADP^+$	C00006	674
NADPH	C00005	671
NAD^+	C00003	640
NADH	C00004	631
ATP	C00002	463
CO_2	C00011	409
Orthophosphate	C00009	381
CoA	C00010	357
ADP	C00008	332
NH_3	C00014	288
Pyrophosphate	C00013	280
UDP	C00015	214

5 Datenvorverarbeitung und Informationsextraktion

Das vorangegangene Kapitel 4 beschäftigte sich mit der Gewinnung der experimentellen und theoretischen Daten. Beide Datensätze sind in ihrer vorliegenden „rohen“ Form für die weiterführende Analyse ungeeignet, sie müssen daher in einem ersten angepassten Schritt vorverarbeitet werden. Für die experimentellen Daten ist hierzu beispielsweise die Plausibilitätskontrolle der Daten, die Detektion von Ausreißern (Kapitel 5.1.1.3) oder auch die Anwendung geeigneter mathematischer Transformationsverfahren (Kapitel 5.1.1.5) zu erwähnen. Bei den theoretischen Daten geht es in erster Linie darum, aus einer komplexen Ausgabedatei relevante Informationen zu extrahieren und damit ebenfalls für nachfolgende Schritte der Analyse nutzbar zu machen (siehe Kapitel 5.2.2). Übergeordnet besitzt dieses Kapitel das Ziel, Informationen aus experimenteller und theoretischer Analyse so reproduzierbar aufzubereiten und standardisiert zu verarbeiten, dass eine Untersuchung in Zusammenschau - mit dem Ziel Auffälligkeiten aufzudecken - erfolgen kann.

5.1 Experimentelle Daten

Wie in Kapitel 4 erwähnt wurden die Metabolitkonzentrationen zeitlich aufgelöst mit Hilfe der GC/MS-Technologie erfasst. Daraus ergibt sich für jeden Metaboliten eine Zeitreihe, welche die zeitliche Veränderung seiner Konzentration beschreibt. Während der Fermentationsexperimente ist *C. glutamicum* darauf angewiesen, das dargebotene Substrat umzuwandeln und zur Energiegewinnung sowie zum Aufbau von Biomasse zu verwenden. Über die Art und Weise, wie der Organismus vermutlich das dargebotene Substrat aufnimmt und in seinem metabolischen Netzwerk weiterverarbeitet, ist schon in zahlreichen Publikationen geforscht

worden. Für einen Überblick ist hierzu insbesondere die zusammenfassende Monographie von Eggeling und Bott aus dem Jahre 2005 zu empfehlen. Aus den zahlreichen Untersuchungen resultiert, dass sich der Stoffwechsel in Abhängigkeit des verfügbaren Ausgangssubstrates grundlegend unterscheidet. Die Betrachtung der gemessenen Konzentrationsverläufe kann darüber Aufschluss geben, wie der Metabolismus unter den gegebenen Fermentationsbedingungen abläuft. Zur Erläuterung seien an dieser Stelle einige Beispiele gegeben. Nimmt beispielsweise ein Metabolit in seiner Konzentration kontinuierlich im Laufe des Fermentationsexperimentes ab, so kann es sich um einen mit dem Ausgangssubstrat in Beziehung stehenden Metaboliten handeln, welcher zum Beispiel zur Energiegewinnung vom Organismus aufgebraucht wird. Bei Metaboliten die in Ihrer Konzentration stetig - bis zum Ende der exponentiellen Wachstumsphase ansteigen, kann es sich um Endprodukte des Stoffwechsels handelt. Verhalten sich beispielsweise zwei Metabolitzeitreihen sehr ähnlich zueinander, so kann es sein, dass sie im metabolischen Netzwerk benachbart und den gleichen übergeordneten regulatorischen Mechanismen unterworfen sind. In diesem Zusammenhang wird deutlich, welche Rolle die theoretischen Daten einnehmen. Sie helfen zu beantworten, ob beispielsweise zwei hoch korrelierte Metabolitpaare im metabolischen Netzwerk benachbart oder weit voneinander entfernt sind. Ist letzteres der Fall könnte dies ein Hinweis auf noch nicht entdeckte regulatorische Zusammenhänge sein (vergleiche hierzu Kapitel 7.4.2.1). Aus oben angerissenen Gründen kommt der Untersuchung der Metabolit-Zeitreihen folglich eine große Bedeutung bei der Klärung zugrundeliegender regulatorischer Prozesse zu. Da sich allerdings die Metaboliten hinsichtlich ihrer Konzentration um mehrere Größenordnungen voneinander unterscheiden können, was oft in keinem direkten Bezug zur biologischen Relevanz steht, ist es unerlässlich, diesen Sachverhalt bei der Vorverarbeitung der Daten ausreichend zu berücksichtigen. Ohne eine Vergleichbarmachung würden die Metaboliten höchster Konzentration die Analyse als auch die Ergebnisse beeinflussen. Eine exemplarische Abbildung zweier Metabolitzeitreihen, welche im Rahmen des Fermentationsexperimentes unter Glucose erfasst wurden, ist in Abbildung 4.6 dargestellt.

Die Konzentrationszeitreihen sind ferner je nach Fermentationsexperiment unterschiedlich lang. Sie erstrecken sich über einen Bereich von 12 Stunden bei der Glucose- Fermentation, bei der *C. glutamicum* vergleichsweise hohe Wachstumsraten erreicht, bis hin zu 28 Stunden bei der Glutamin-Fermentation, bei der

C. glutamicum sehr langsam wächst. Aus messtechnisch bedingten Gründen wurden die Konzentrationen hierbei in einem Abstand von einer Stunde bestimmt. Auch die Anzahl der messtechnisch erfassbaren Metaboliten ist je nach Fermentation unterschiedlich. So konnten beispielsweise bei der Fermentation mit Glutamin 138 individuelle Metaboliten detektiert werden, während es bei der Fructose Fermentation 172 waren. Im Vergleich der Fermentationen können nicht detektierte Metaboliten unter Umständen ein Hinweis darauf sein, dass der fehlende Metabolit unter den gegebenen Umweltbedingungen keine Rolle spielt. Im unverarbeiteten Zustand bestehen die Konzentrationsverläufe aus so genannten „Pseudo-Amount"-Werten, die bei der Integration Peakflächen im Schritt der Metabolitquantifizierung (siehe Kapitel 4.2.3) entstehen. In den nachfolgenden Kapiteln wird dargelegt, welchen Vorverarbeitungsschritten die Rohdaten bis hin zur Analyse unterzogen werden.

5.1.1 Vorverarbeitung der experimentellen Daten

Jeder tiefergehenden mathematisch-statistischen Analyse oder Klassifikationsfragestellung sollte der Schritt einer ausführlichen Datenvorverarbeitung vorangehen. Denn Rohdaten sind im Allgemeinen fehlerbehaftet, das heißt, sie weisen oft Auffälligkeiten oder Artefakte auf, die vor allem bei der Anwendung multivariater statistischer Verfahren oder maschineller Lernsysteme genauer untersucht und gegebenenfalls eliminiert werden müssen. So können etwaige in den Daten vorhandene „Fehler" unter Umständen die Aussage der Datenstrukturanalyse verzerren, die Präzision von Vorhersagen beeinflussen oder diese sogar gänzlich unmöglich machen. Dies bedeutet, die wissenschaftliche Analyse eines Datensatzes muss die Fehlerproblematik berücksichtigen, denn ohne Kenntnis der Fehlerstruktur ist die Aussage wertlos.

In der Mehrzahl der Fälle sind die Daten, mit denen man arbeitet nicht von einem selbst erhoben worden - über ihre Entstehung ist oft so gut wie nichts bekannt. Ist dies der Fall, muss eine Datenvorverarbeitung umso ausführlicher durchgeführt werden, denn der einzige „Zeuge" für etwaige, während einer Analysereihe aufgetretene, Schwierigkeiten sind die Messungen selbst. Besser ist es natürlich, wenn eine ausführliche Labordokumentation oder der Ansprechpartner, der die Daten erhoben hat, für Rückfragen zur Verfügung steht. Im Rahmen dieser Arbeit war dies glücklicherweise der Fall. Jegliche Zusatzinformation, die zu

einem besseren Verständnis der Daten führt ist, daher von unschätzbarem Wert, vor allem dann, wenn es sich um eine komplexe Datenstrukturanalyse oder ein schwieriges Vorhersageproblem handelt.

Die Verwendung von metabolischen Daten im Kontext dieser Arbeit stellt eine zusätzliche Herausforderung für die Datenvorverarbeitung dar, worauf in den nächsten Kapiteln gesondert eingegangen wird. Um dies zu berücksichtigen, wurde eine auf die Verwendung metabolischer Daten abgestimmte Prozedur der Datenvorverarbeitung konzipiert und angewandt. Sie gliedert sich in mehrere aufeinander folgende Schritte, welche schematisch in nachfolgender Abbildung 5.1 dargestellt sind.

5.1.1.1 Einlesen der Rohdaten

Nach der Metabolitquantifizierung mit Xcalibur (siehe Kapitel 4.2.3) liegen die Daten in Form einer Textdatei vor, bei der die Konzentrationen in Zeilenform hinterlegt sind. Die Anzahl von Spalten entspricht hierbei der Anzahl von Messpunkten während des Fermentationsexperimentes wobei die Wiederholungsmessungen eines Zeitpunktes als solche gekennzeichnet sind. Der Anzahl von Zeilen in dieser Matrix entspricht der Anzahl der detektierten Metaboliten. Für dieses Ausgangsformat wurde eine Einleseroutine in MATLAB geschrieben, mit dem Ziel die Daten aller Fermentatiosexperimente einheitlich für die weitere Verarbeitung zugänglich zu machen.

5.1.1.2 Auswahl von Stereoisomeren

Wie bereits im Kapitel 4.2.3 angerissen, ist die Massenspektroskopie nicht in der Lage, zwischen Stereoisomeren zu unterschieden. Als Stereoisomere versteht man solche chemischen Verbindungen, die sich nicht in ihrer atomaren Zusammensetzung unterscheiden, aber durchaus eine unterschiedliche räumliche Orientierung wie zum Beispiel eine spiegelverkehrte Anordung von Seitengruppen etc. besitzen können. Aus diesem Grunde gibt die Detektionssoftware AMDIS alle für einen Treffer gefundenen Stereoisomere aus, selbst wenn nur einer davon tatsächlich nach aktuellem Stand des Wissens in *C. glutamicum* vorkommt. Hieraus resultiert, dass aufgrund der Isomere in dem oben beschriebenen Datenformat redundante Zeilen existieren. Da allerdings eine solche Mehrfacheintragung bei der mathematischen Datenvorverarbeitung eine Übergewichtung hervorrufen würde, müssen

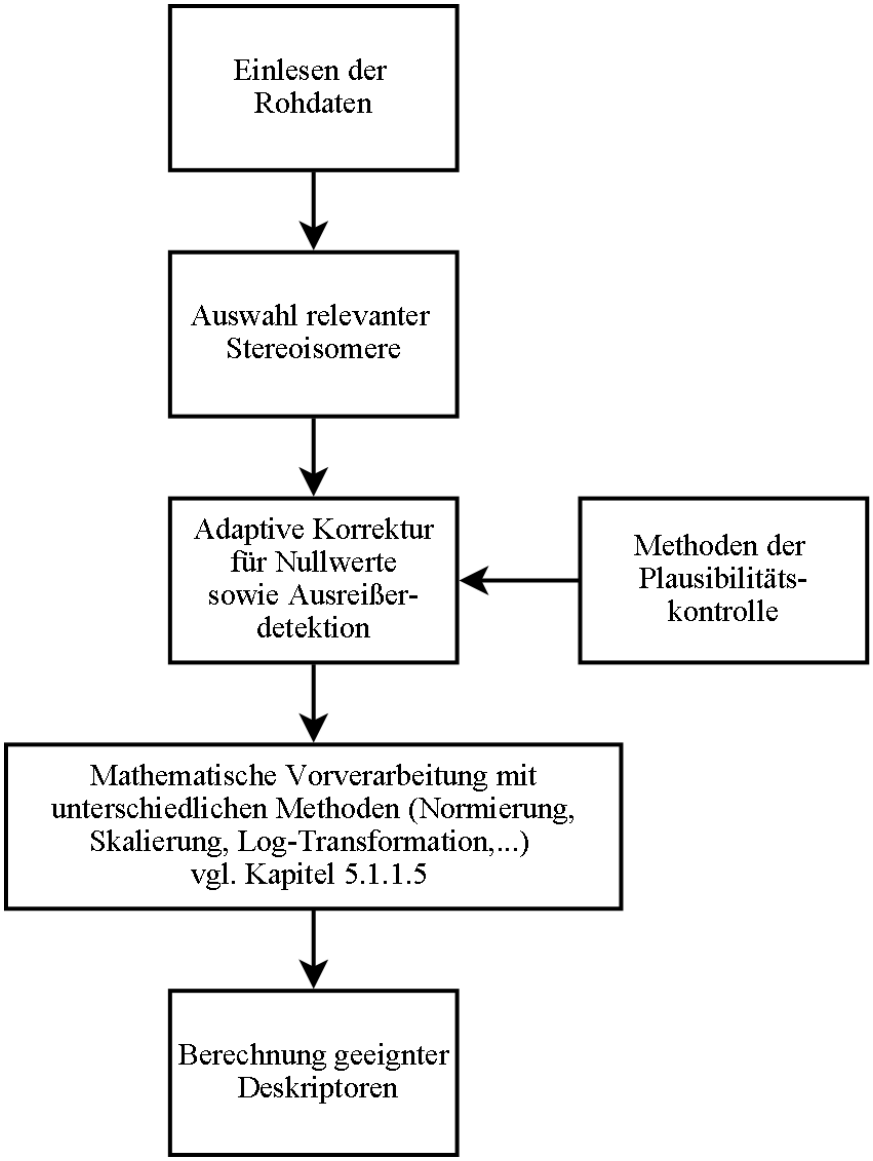

Abbildung 5.1: Schematische Darstellung der Datenvorverarbeitung auf den Zeitreihen der Metabolitkonzentration

die redundanten Zeilen nach biologischen Gesichtspunkten entfernt werden. Im Allgemeinen wurden jene Isomere von Metaboliten verworfen, die im Stoffwechsel von *C. glutamicum* nicht von physiologischer Bedeutung sind. Hierzu gehören: alle Aminosäuren mit D-Konfiguration sowie alle Monosaccharide mit L-Konfiguration bis auf L-Arabinose. In uneindeutigen Fällen, bei denen in denen die Bedeutung nicht klar abschätzbar war, wurden jene Isomere ausgewählt, die organismenspezifisch an den meisten Reaktionen beteiligt sind. Auch die Informationen hierzu wurden aus der Datenbank KEGG entnommen. Um Reproduzierbarkeit zu gewährleisten, wurde in den Datensätzen vermerkt, welche Stereoisomere eliminiert wurden.

5.1.1.3 Detektion von Ausreißern

Selbst wenn unter extrem reproduzierbaren Randbedingungen gearbeitet und gemessen wurde, kann es vorkommen, dass sich einzelne Daten des Kollektivs gänzlich anders verhalten. Hierbei spricht man von sogenannten Ausreißern (outliers). Die Gründe für das Zustandekommen von Ausreißern können mannigfaltig sein, ebenso wie die Gestalt der Ausreißer. Aus diesem Grunde ist es umso wichtiger, vor einer weiteren Prozessierung der Daten Ausreißer zuverlässig zu identifizieren und von gegebenenfalls zu korrigieren. Im Rahmen dieser Arbeit wurde eine Ausreißerdetektion mit Hilfe des „Grubbs“-Algorithmus durchgeführt (Massart et al., 1997). Als Ausreißer detektierte Datenpunkte wurden unter Hinzuziehung von Expertenwissen korrigiert und in seltenen Fällen auch von der weiteren Analyse ausgeschlossen.

5.1.1.4 Adaptive Korrektur für fehlende Werte

Ähnlich wie die Ausreißer stellen auch fehlende Werte ein Problem für die nachfolgende Analyse dar. So kommt es bei den Triplikaten beispielsweise vor, dass einzelne Messungen Nullwerte besitzen. Um mit dieser Problematik umzugehen, wurden eine Reihe von Regeln angewandt, wie sie in Anlehnung an Dr. Silke Schraders Tool zur Verarbeitung metabolischer Daten (CUMETA) implementiert sind. Dieses Vorgehen berücksichtigt Eigenheiten, wie sie bei der Metabolitquantifizierung auftreten können. Da die Quantifizierung graphisch durch Integration der Peakflächen erfolgt, kann es sein, dass extrem kleine Metabolitpeaks nur äußerst schwierig zu integrieren sind. Wurde ein Metabolit nicht detektiert, so findet

sich ein Nullwert in der Datei. Wie mit einem Nullwert umzugehen ist, ist folglich davon abhängig, welchen Wert die beiden verbliebenen Wiederholungsmessungen für den gleichen Messzeitpunkt besitzen. Im Einzelnen finden bei unterschiedlichen Szenarien vorab definierte Regeln Anwendung, wobei ein Schwellenwert Berücksichtigung findet. Diese sind nachfolgend angegeben.

- Befindet sich in einem Triplikat ein Nullwert, während sich die restlichen oberhalb des vordefinierten Schwellenwert befinden, so wird der Nullwert aus dem Mittelwert der beiden verbleibenden Messungen ersetzt. Für die Berechnung eines Mittelwertes aus den Wiederholungsmessungen werden alle Messungen verwendet. (Fall A in Tabelle 5.1)

- Befindet sich in einem Triplikat ein Nullwert, während die beiden anderen Messwerte unterhalb des vordefinierten Schwellenwertes liegen, so wird der Nullwert durch 1 ersetzt. Für die Mittelwertsberechung zu einem gegebenen Zeitpunkt werden alle Wiederholungsmessungen verwendet. (Fall B in Tabelle 5.1)

- Befindet sich in einem Triplikat ein Nullwert, während einer der beiden Messwerte unterhalb des vordefinierten Schwellenwertes liegt, so wird der Nullwert durch 1 ersetzt, aber für die Mittelwertsberechnung aus allen Wiederholungsmessungen nicht berücksichtigt. (Fall C in Tabelle 5.1)

- Befinden sich in einem Triplikat zwei Nullwerte, während der verbliebene Messwert unterhalb des vordefinierten Schwellenwertes liegt, so werden die Nullwerte durch 1 ersetzt und bei der nachfolgenden Mittelwertsberechnung einbezogen. (Fall D in Tabelle 5.1)

- Enthält ein Triplikat nur Nullwerte, werden alle drei Werte auf den Wert 1 gesetzt und zur Mittelwertsberechnung einbezogen. (Fall E in Tabelle 5.1)

Die Mittelwertsberechnung ist von besonderer Bedeutung, da für die weitere Analyse ausschließlich die aus den Wiederholungsmessungen abgeleiteten Mittelwerte genutzt werden. Einzige Ausnahme stellt lediglich die Berechnung der paarweisen Gleichläufigkeit dar, die zusätzlich die Information des Triplikates berücksichtigt (vergleiche Kapitel 5.1.3.3). Aus Vorversuchen konnte abgeleitet

werden, dass ein Schwellenwert von 1000 für die zugrunde liegenden Daten vernünftige Ergebnisse liefert. Der Ersetzung der Nullwerte durch den Wert 1 dient dazu, im nachfolgenden Schritt die Berechnung des Logarithmus zu ermöglichen.

Tabelle 5.1: Beispielhafte adaptive Korrektur für Nullwerte bei dreifachen Konzentrationswerten und einem Schwellenwert von 1000

	Originale Daten			Verarbeitete Daten			
Triplikat	1	2	3	1	2	3	Mittelwert
cA	0	5000	7000	6000	5000	7000	6000
cB	0	500	700	1	500	700	400,3
cC	0	5000	700	1	5000	700	2850
cD	0	0	600	1	1	600	200,6
cE	0	0	0	1	1	1	1

Diese Überlegungen lehnen sich dem Vorgehen der Plausibilitätskontrolle an. Hierunter versteht man einen mathematisch-basierten Ansatz, um ohne Zusatzinformationen zu überprüfen, ob ein vorliegender Datensatz plausibel ist. Vereinfacht ausgedrückt bedeutet dies, festzustellen, ob eine Messung richtig oder falsch ist; wenn sie falsch ist: zu überprüfen, ob sie korrigierbar ist, und wenn ja: wie.

5.1.1.5 Mathematische Vorverarbeitung mit unterschiedlichen Methoden

Bei der Verwendung von Daten aus der Metabolomforschung gibt es charakteristische Besonderheiten, die für die Datenvorverarbeitung eine Herausforderung darstellen und deshalb zu berücksichtigen sind. Wie bereits erwähnt, kann es durchaus vorkommen, dass Metaboliten sich in ihren experimentell erfassten Konzentationensverhältnissen um mehrere Größenordnungen voneinander unterscheiden, wobei diese überproportionalen Konzentrationsunterschiede in keiner direkten Beziehung zur biologische Relevanz stehen (van den Berg et al., 2006). Diese Eigenschaften gilt es durch geeignete mathematische Vorverarbeitungsstrategien zu berücksichtigen. Analog sind auch die Autoren oben genannter Veröffentlichung vorgegangen. Sie haben unterschiedliche Vorverarbeitungstrategien auf einem Beispieldatensatz durchgeführt und konnten dabei feststellen, dass die Wahl des Vorverarbeitung die Aussage der anschließenden statistischen Analyse (wobei eine Hauptkomponentenanalyse verwendet wurde) zum Teil erheblich verändert. Da in dieser Arbeit der Einfluss der mathematischen Vorverarbeitung

auf die spätere Analyse getestet werden soll, werden insgesamt an dieser Stelle vier verschiedene mathematische Ansätze angewandt. Grundlage sind jeweils die, wie im adaptiven Verfahren in Kapitel 5.1.1.3 und 5.1.1.4 beschrieben, aus den Triplikaten hervorgegangenen Mittelwerte für jeden Zeitpunkt der betrachteten Zeitreihe. Bei der mathematischen Vorverarbeitung handelt es sich um folgende Ansätze:

- Der dekadische Logarithmus der Konzentrationswerte wird berechnet. Dieses Vorgehen dient der Korrektur von Heteroskedastizität, also etwaiger in den Daten vorhandener intrinsicher Verzerrung und ist im Bereich der Metabolomforschung weit verbreitet.
- Bei der Medianzentrierung werden alle Datenpunkte einer Zeitreihe durch den Median der gesamten Zeitreihe dividiert. Dies hat zur Konsequenz, dass etwaige Offsets in den Zeitreihen eliminiert werden.
- Autoskalierung: Die Autoskalierung erfolgt, indem von jedem Zeitpunkt t der Mittelwert der gesamten Zeitreihe subtrahiert wird. Die Differenz wird anschließend durch die Standardabweichung der Zeitreihe dividiert. Dieses Vorgehen bewirkt, dass alle Metaboliten die gleiche Wichtigkeit erlangen.
- Vektornormierung: Hierbei wird der Mittelwert der Zeitreihe berechnet. Die Zeitreihe wird anschließend um den Betrag jenes Wertes subtrahiert. Anschließend erfolgt die Berechnung der Vektorlänge der Zeitreihe. Abschließend wird das Datenkollektiv durch diesen Wert dividiert.

Im Batch-Verfahren wurden die Vorverarbeitungsschritte einzeln, aber auch in definierten Kombinationen (beispielsweise Logarithmierung der Daten mit anschließender Medianzentrierung) in paralleler Form durchgeführt. Ziel war es hier, den Einfluss unterschiedlicher Vorverarbeitungsmethoden auf die Berechnung der Deskriptoren und die anschließende gemeinsame Analyse experimenteller und theoretischer Daten zu bestimmen und die optimale Vorverarbeitungsstrategie für die zu beantwortenden Fragestellungen auszuwählen.

5.1.2 Definition eines Ähnlichkeitsbegriffs auf experimentellen Daten

Bei der Beschreibung der Zeitreiheneigenschaften werden zum einen individuelle Zeitreihen betrachtet, als auch paarweise Vergleiche von Zeitreihen durchgeführt. Während zur Charakterisierung einzelner Zeitreihen in erster Linie beschreibende Größen Verwendung finden, ist beim paarweisen Vergleich eine vorherige Definition eines Ähnlichkeitsbegriffes unabdingbar. Bei den erfassten Zeitreihen der Metabolitkonzentrationen handelt es sich um Informationen, die einen stark prozessbezogenen Charakter haben. So verändert sich der zeitliche Verlauf nicht zufällig, sondern ist von regulatorischen Mechanismen abhängig, welche zu den jeweiligen Zeitpunkten stattfinden. Die vorhandenen Zeitreihen stellen folglich das Resultat von komplexen regulatorischen Prozessen, welche entlang der Wachstumskurve ineinandergreifen, dar. Der Form des Konzentrationsverlaufes wird daher besondere Bedeutung beigemessen, was sich bei der Auswahl geeigneter Deskriptoren im nachfolgenden Kapitel niederschlägt.

5.1.3 Auswahl geeigneter Deskriptoren

Diese prozessabhängige Betrachtungsweise muss die Auswahl geeigneter beschreibender Größen für den paarweisen Vergleich zwischen zwei Metabolitkonzentrationen nach sich ziehen. Neben der Fragestellung ob die gemessenen Metabolitkonzentrationen in einem statistischen Zusammenhang zueinander stehen (u.a. berechnet durch den Korrelationskoeffizienten) muss auch die Betrachtung der Formähnlichkeit zwischen zwei Konzentrationsverläufen Berücksichtigung finden (u.a. berechnet durch die Winkelähnlichkeit und die Gleichläufigkeit).

Sämtliche, die paarweise Ähnlichkeit zweier Konzentrationszeitreihen, beschreibenden Größen werden in den nächsten Kapiteln Deskriptoren genannt. Für die Berechnung der Deskriptoren auf den experimentellen Daten wurden eine Reihe von unterschiedlichen Ansätzen verwendet, die nachfolgend beschrieben werden. Die betrachteten Konzentrationsverläufe zweier Metaboliten werden hierbei als x und y bezeichnet.

5.1.3.1 Korrelationsberechnung

Die Berechnung von Korrelationen wird bereits im Bereich der Metabolomforschung zur Untersuchung statistischer Zusammenhänge zwischen Metaboliten verwendet (Steuer et al., 2003). Im Rahmen dieser Arbeit wurden der Pearsonsche Korrelationskoeffizient, (Formel 5.1) sowie der Spearmansche Rangkorrelationkoeffizient (Formel 5.2) zur Berechnung der Ähnlichkeit zweier Konzentrationszeitreihen verwendet.

$$d_{Pearson} = \frac{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right)}{\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}\right)} \tag{5.1}$$

$$d_{Spearman} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \text{wobei} \quad d_i = Rang(x_i) - Rang(y_i) \tag{5.2}$$

Die Korrelationsanalyse liefert einen Hinweis darauf, ob zwei Datenkollektive einen statistisch auffälligen Zusammenhang aufweisen. Die Korrelationsberechnung wird auf den vorverarbeiteten Daten, sowohl zusätzlich auf der ersten Ableitung derselben durchgeführt.

5.1.3.2 Winkelähnlichkeit

Die Winkelähnlichkeit (Formel 5.3) erlaubt eine Aussage darüber, wie ähnlich zwei Datenreihen hinsichtlich Ihrer Form sind. Die Winkelähnlichkeit ist unabhängig von der Intensität der zu vergleichenden Datenreihen und findet im Bereich der Spektroskopie seit Anfang der 1990er Jahre Anwendung (Kruse et al., 1993). Hierbei werden die Konzentrationszeitreihen als n-dimensionale Vektoren betrachtet, wobei die Anzahl der Dimensionen der Anzahl der Zeitpunkte entspricht. Zwischen den beiden Vektoren wird, vom Ursprung des Koordinatensystems aus gesehen, ein Winkel berechnet. Dieses Vorgehen hat den Vorteil, dass es in erster Linie die Formähnlichkeit der Profile zueinander charakterisiert. Eine Winkelähnlichkeit von 0° würde einer exakten Deckungsähnlichkeit entsprechen, während einem Wert von 180° Datenkollektive mit invertierten Vorzeichen entsprechen. Natürlich wird bei der Anwendung auf die Konzentrationszeitreihen nicht der ganze zur Verfügung stehende Wertebereich abgebildet. So zeigten Vorversuche

auf diesen Daten, dass eine Spanne der Winkelähnlichkeit von knapp über 2° bis hin zu ca. 120° abgebildet wird.

$$d_{Winkelaehnlichkeit} = \cos^{-1}\left(\frac{\sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}} (\sum_{i=1}^{n} y_i^2)^{\frac{1}{2}}}\right) \tag{5.3}$$

Wie auch die Korrelationsberechnung wird die Berechnung der Winkelähnlichkeit auf den vorverarbeiteten Daten sowohl deren erster Ableitung durchgeführt.

5.1.3.3 Gleichläufigkeit

Die Gleichläufigkeit ist ein aus der Dendrochronologie hervorgegangener Vorzeichentest, der in dieser Form von Schweingruber entwickelt wurde (Schweingruber, 1983). Im Rahmen dieser Arbeit wird eine leicht abgewandelte Form genutzt. Zur Berechnung werden die fertig vorverarbeiteten Daten verwendet. Im Einzelnen wird zuerst für jeden Zeitpunkt t der Konzentrationszeitreihe der Mittelwert und die Standardabweichung aus den Dreifachmessungen berechnet. Anschließend wird die Differenz eines jeden Mittelwertes zum nächsten berechnet. Ist die Differenz zweier benachbarter Mittelwerte kleiner als die beiden Standardabweichungen der jeweiligen Punkte, so wird von Beginn an die Veränderung des Kurvenverlaufs als nicht signifikant angesehen, d.h. die Differenz gleich Null gesetzt (Formel 5.4) .

$$\text{Wenn:} \quad (\overline{x}_{t+1} - \overline{x}_t) < max(\sigma_{x(t)} | \sigma_{x(t+1)}) \quad \text{dann:} \quad (\overline{x}_{t+1} - \overline{x}_t) = 0 \tag{5.4}$$

Für den paarweisen Kurvenvergleich gilt nun folgendes: Sind für beide Kurven x und y die Differenzen zweier benachbarter Punkte $>$ Null, so wird für die Gleichläufigkeit der Wert 1 gegeben. Sind beide Werte $<$ Null oder gleich null (bzw. innerhalb der Standardabweichung), so wird ebenfalls der Wert 1 gegeben. In diesem Falle verhalten sich beide Kurven gleichartig. Verändert sich nur eine Kurve, während bei der anderen die Differenz gleich Null ist, so wird der Wert 0,5 vergeben. Verändern sich beide Kurven gegenläufig so wird der Wert Null vergeben. Nachfolgende Tabelle 5.2 verdeutlicht das Bewertungsschema:

Die Werte für jeden beobachteten Zeitpunkt werden aufaddiert und um die

Tabelle 5.2: Bewertungsschema zur Bestimmung der Gleichläufigkeit (G_t) zwischen zwei Datenreihen zum Zeitpunkt t.

		$(x_{t+1}$ -$x_t)$		
	G_t	>0	=0	<0
$(y_{t+1}$ -$y_t)$	>0	1	0,5	0
	0	0,5	1	0,5
	<0	0	0,5	1

Länge der Datenreihe dividiert. Ein Wert um 1 zeigt eine hohe Konvergenz, ein Wert um Null eine hohe Divergenz an.

$$d_{Gleichlaeufigkeit} = \frac{1}{t-1} \sum_{t=1}^{t-1} G_t \tag{5.5}$$

Aus diesem Vorgehen resultiert, dass die Gleichläufigkeit als Deskriptor nur diskrete Wertestufen annimmt und ungleich zu Korrelation und Winkelmaß keinen kontiniuerlichen Wertebereich abdeckt.

5.1.3.4 Log-10 Ratios

Es werden auch hier die vorverarbeiteten Daten verwendet. Zunächst wird der Mittelwert aus den Triplikaten gebildet. Von diesen Mittelwerten ausgehend wird anschließend für jeden Zeitpunkt der Reihe das Verhältnis der Metaboliten x und y zueinander berechnet. Für jedes Verhältnis wird anschließend der Logarithmus zur Basis 10 gebildet. Für den paarweisen Vergleich von zwei Metaboliten wird die Standardabweichung in % vom Mittelwert für die Verhältniszeitreihe ermittelt (Formel 5.6).

$$d_{logratio} = \frac{\sigma(a_{i\ldots n})}{\overline{a}_{i\ldots n}} \times 100\ [\%] \qquad \text{wobei} \quad a_{i\ldots n} = \log\left(\frac{x_{i\ldots n}}{y_{i\ldots n}}\right) \tag{5.6}$$

5.1.3.5 Sensitivität

Die Sensitivität ist in diesem Zusammenhang als ein Maß zur Charakterisierung der Variabilität einer Datenreihe zu verstehen. Die lokale Sensitivität S_i ist als die Differenz eines Datenwertes zu seinem Nachfolger dividiert durch den Mittelwert der beiden Werte definiert. Zur Berechnung der so genannten globalen Sensitivität über eine gesamte Datenreihe bildet man das arithmetische Mittel über alle lokalen Sensitivitäten. Damit lässt sich eine einzelne Zeitreihe charakterisieren.

$$d_{sens} = \frac{\sum_{i=2}^{n} |S_i|}{n-1} \qquad \text{wobei} \quad S_{i+1} = 2\frac{(x_{i+1} - x_i)}{(x_{i+1} + xi)} \tag{5.7}$$

Zum Vergleich zweier Datenreihen lässt sich der Mittelwert des Betrages der beiden globalen Sensitivitäten berechnen. Je kleiner dieser Wert, desto geringer die Schwankungsbreite der verglichenen Metabolitprofile.

5.1.3.6 Mutual Information

Die „Mutual Information“ ist ein Entropiemaß, welches beschreibt, wie viel gegenseitige Information zwei Datenreihen zueinander enthalten. Hierzu werden die Wahrscheinlichkeitsverteilungen der beiden Variablen (in diesem Fall Metabolitprofile) betrachtet. Dieses Prinzip wurde bereits als Ähnlichkeitsmaß in der Analyse von cDNA Microarrays als auch in der Metabolomanalyse angewendet (Steuer et al., 2002). In Rahmen dieser Arbeit wurde die verallgemeinerte Berechnungsform verwendet (Formel 5.8).

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)} \tag{5.8}$$

Wie bereits angerissen, ist für die integrative Analyse experimenteller und theoretischer Daten die Schnittmenge zwischen beiden Datensätzen maßgebend. Wie im späteren Verlauf dieser Arbeit gezeigt wird, wurden die paarweisen Deskriptoren zwischen jenen Metabolitpaaren berechnet, zwischen denen laut theoretischer Betrachtungsweise ein metabolischer Pfad existiert.

5.2 Theoretische Daten

Unter dem Begriff „theoretische“ Daten werden alle jene Informationen zusammengefasst, die aus der Rekonstruktion der metabolischen Netzwerke, sowie der anschließenden Modellierung der Stoffwechselwege abgeleitet worden sind.

Wie in Kapitel 4 beschrieben, wurden analog zu den experimentellen Untersuchungen, auch die Netzwerkmodellierungen in mehreren Versuchsreihen durchgeführt. Insbesondere wurden dabei vier verschiedene Reaktionsnetzwerke analysiert, welche zum einen aus der Bielefelder- (CGB) und Kyowa Hakko- (CGL) Annotation von *C. glutamicum*, als auch der im Rahmen dieser Arbeit durchgeführten Genomannotation, entstammen (VGL1 und VGL2).

5.2.1 Vorverarbeitung der theoretischen Daten

Das Pathway Hunter Tool (PHT) liefert, wie in Kapitel 4.3.3 beschrieben, für die Modellierung eines metabolischen Netzwerkes eine Fülle von Informationen, die in einer einzigen großen Textdatei abgelegt werden. Um die gewünschten Informationen zu extrahieren, wurde in MATLAB ein Programm implementiert, das den von PHT gelieferten Output hinsichtlich relevanter Informationen untersucht und die Informationen in verwertbarer Form extrahiert. Das Programm untersucht die Ausgabedatei anhand von Schlüsselworten und Textmustern, formatiert diese um und legt sämtliche Informationen in einer leicht zugänglichen Datenmatrix ab. Durch das Konvertieren verringert sich der Speicherbedarf zudem erheblich und die gewonnenen Informationen können schneller und einfacher weiterprozessiert werden.

5.2.2 Ableitung geeigneter Deskriptoren auf den theoretischen Daten

Um geeignete Deskriptoren aus dem von dem Pathway Hunter Tool gelieferten Output ableiten zu können, muss zuerst die Funktionsweise des Programms und die Struktur der Ausgabedatei verstanden werden. Das PHT berechnet primär den so genannten kürzesten Pfad (Shortest Path) einer Reaktionskette zwischen zwei gegebenen Metaboliten innerhalb eines vordefinierten Reaktionsnetzwerkes wobei Ansätze der Graphentheorie Verwendung finden. Das der Modellierung zugrundeliegende Reaktionsnetzwerk kann aus vordefinierten Einträgen gewählt werden,

oder wie in Kapitel 4.3.1 beschrieben, aus einer eigenen Annotation erzeugt werden.

Um den kürzesten Pfad zu finden, benötigt das PHT die Eingabe von zwei Metaboliten, welche eindeutig über ihre sogenannte C-Nummer (KEGG-Nomenklatur) zu identifizieren sind. Der erstere fungiert als Edukt von dem die Reaktionskette ausgeht, der zweite als Produkt bei dem die Reaktionskette endet. Durch ein Vertauschen der beiden Metaboliten kann folglich also auch die Richtungsabhängigkeit bei der Suche nach metabolischen Pfaden analysiert werden. Dies resultiert in unterschiedlichen Ergebnissen, da bei weitem nicht alle Reaktionen des Reaktionsnetzwerkes reversibel ablaufen können. Im Rahmen dieser Arbeit wurde das PHT mit Hilfe eines Batch-Skriptes für eine Liste von 123 Metaboliten, welche experimentell in allen betrachteten Fermentationen mit *C. glutamicum* gleichermaßen gefunden werden konnten, gestartet. In dieser Einstellung wurde für jede theoretisch denkbare Kombination eine Suche nach dem kürzesten Stoffwechselweg, der beide Metaboliten miteinander verbindet, vorgenommen. Weiterführende Hinweise zum „Shortest Path-“ Algorithmus, der diese Berechnung ermöglicht, findet sich in Kapitel 5.2.3.1. Logischerweise wurden solche Kombinationen, bei denen Start- und Endmetabolit identisch waren, eliminiert. Für die Liste von 123 Metaboliten ergeben sich folglich 15006 mögliche Kombinationen, für die eine Suche nach Stoffwechselwegen durchgeführt wurde.

Für bei weitem nicht alle Metabolitkombinationen konnte tatsächlich ein verbindender Pfad im metabolischen Netzwerk gefunden werden, wobei eine Abhängigkeit von der Komplexität des betrachteten Reaktionsnetzwerkes sowie insbesondere der Anzahl vorhandener Enzyme festzustellen ist. Studien wie die von Ma und Zeng (2003a) konnten belegen, dass in zahlreichen betrachteten Organismen bei weitem nicht alle Metaboliten durch Reaktionswege ineinander überführt werden können. Vereinfacht ausgedrückt bedeutet dies, dass metabolische Netzwerke - nach aktuellem Stand des Wissens - nicht vollständig konnektiert sind. In nachfolgender Tabelle 5.3 ist angegeben, für wie viele der 15006 getesteten Metabolitkombinationen tatsächlich ein oder mehrere Pfade gefunden werden konnten. Für den Fall, dass mehrere alternative (gleich kurze) metabolische Pfade für eine Kombination von zwei Metaboliten existieren, so werden diese bei der Modellierung ausgegeben aber nicht mehrfach gezählt. Detaillierter wird über diese Ergebnisse und die zugrundeliegenden Ursachen im Ergebnisteil (Kapitel 7.2) eingangen.

Zur Illustration der Thematik der Findung metabolischer Pfade, soll an dieser

Tabelle 5.3: Tabellarischer Vergleich gültiger Metabolitkombinationen, für die in Abhängigkeit der betrachteten Reaktionsnetzwerke, metabolische Pfade gefunden werden konnten.

Parameter	CGB	CGL	VGL1	VGL2
KEGG Mapping	2559	2543	3725	3901
CUBIC Mapping	3903	3862	4682	4902

Stelle Abbildung 5.2 dienen. Sie zeigt eine beispielhafte schematische Darstellung jener metabolischen Pfade welche ausgehend vom Metaboliten beta-D-Glucose 6-phosphate (C01172) zu Pyruvate (C00022) für die Untersuchung mit dem Pathway Hunter Tool gefunden werden konnten. Der ermittelte Pfad gehört zum Stoffwechselweg der Glykolyse, bei dem Kohlenhydrate wie Glucose unter Energiegewinnung zu Pyruvate abgebaut werden. Bei der theoretischen Untersuchung werden zwei gangbare Pfade ermittelt, die jeweils 8 Reaktionsschritte lang sind. Der erste Pfad nimmt von beta-D-Fructose 6-Phosphate (C05345) unter Verwendung des Enzyme 6-Phosphofructokinase (EC 2.7.1.11) und Fructose bisphosphat Aldolase (EC 4.1.2.13) den Weg zu D-Glyceral- dehyde 3-phosphate (C00118). Der alternative Pfad wird zweifach durch das Enzym Transketolase (EC 2.2.1.1) katalysiert und führt über einen Metaboliten des Pentose-Phosphat-Weges, namentlich D-Xylulose 5-Phosphate (C00231). Wie sichtbar wird, besteht ein metabolischer Pfad aus mindestens einem, meistems aber mehreren aufeinander folgenden Reaktionsschritten, die - enzymatisch katalysiert - das Edukt in das Produkt umsetzen. Der kürzeste denkbare metabolische Pfad besitzt folglich die Länge 1, das bedeutet, das Edukt kann durch einen einzigen Reaktionsschritt in das Produkt umgesetzt werden. Wenn man sich eine Häufigkeitsverteilung über die gefundenen metabolischen Pfade und deren Schrittlänge anschaut, so zeigt sich im Falle von *C. glutamicum* für die beschriebenen 123 Metaboliten und deren Kombinationen eine linksschiefe Verteilung mit einem Maximum bei ca. 7 Reaktionsschritten.

Um möglichst viel an Information aus der Ausgabedatei zu extrahieren und für die nachfolgende Analyse nutzbar zu machen, wurden Skripte entwickelt, die auf der kleinsten funktionellen Einheit, dem einzelnen Reaktionsschritt detaillierte Informationen sammeln. Die Informationen der zu einem Pfad gehörigen Reaktionsschritte wurden durch mathematische Operationen miteinander verknüpft. Das übernächste Kapitel 5.2.4 beschäftigt sich mit der Fusionierung der zu ei-

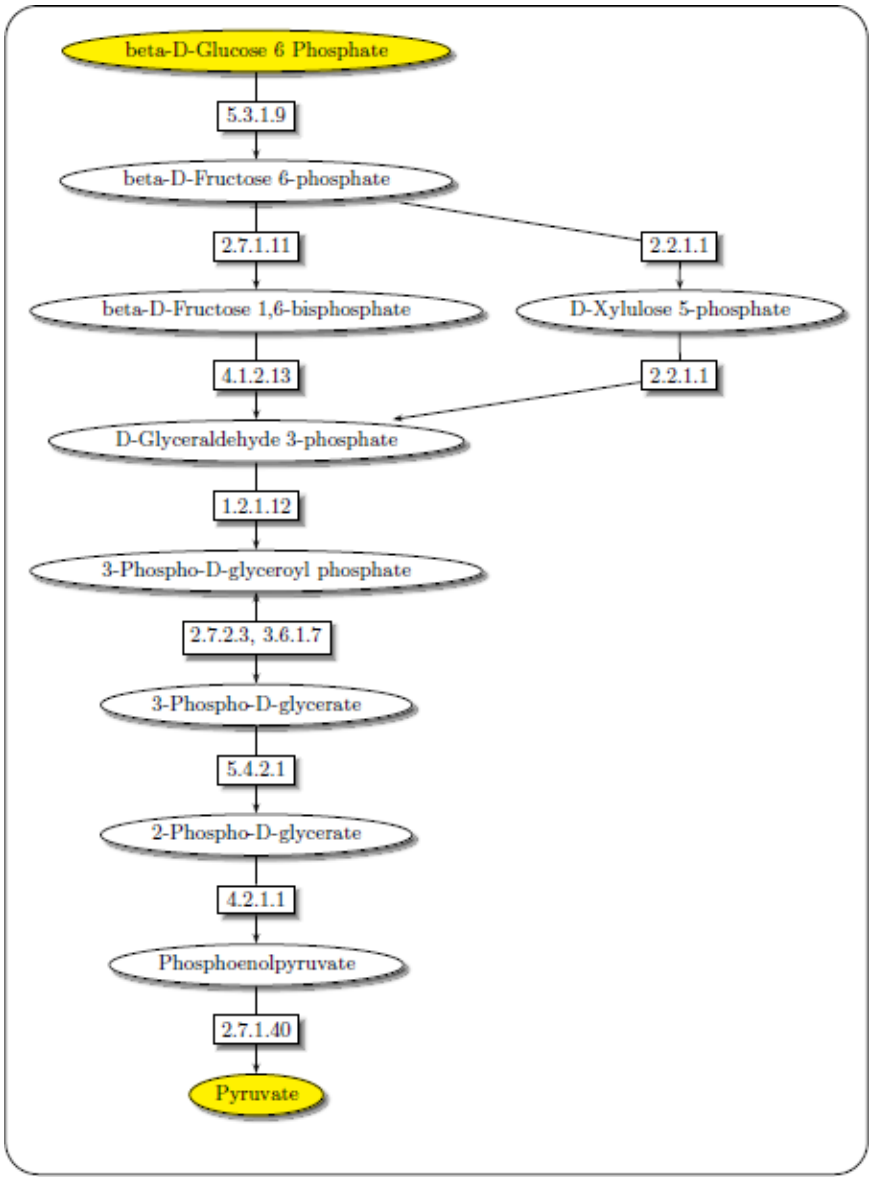

Abbildung 5.2: Beispiel für mit dem Pathway Hunter Tool ermittelte Pfade vom Metaboliten beta-D Glucose 6-Phosphate (C01172) ausgehend zu Pyruvate (C00022). Schematische Darstellung der Reaktionsrichtungen, sowie der katalysierenden Enzyme.

nem metabolischen Pfad gehörenden Deskriptoren auf der Ebene von einzelnen Reaktionsschritten.

5.2.3 Direkte theoretische Deskriptoren

Bei den direkten Deskriptoren handelt es sich um solche, die den metabolischen Pfad als solchen beschreiben und ohne Integration von Detailinformationen der einzelnen Reaktionsschritte abzuleiten sind. Hierzu gehören in erster Linie: der kürzeste Reaktionsabstand zwischen zwei Metaboliten, die Anzahl gleich kurzer Pfade zwischen zwei Metaboliten, sowie eine eigens erstellte Größe, welche beschreibt, ob und wie weit der betrachtete Pfad seinen Weg über den Zitratzyklus nimmt. Die erwähnten Deskriptoren werden in den folgenden Unterkapiteln weitergehend erläutert.

5.2.3.1 Kürzester Pfad zwischen zwei Metaboliten

Der kürzeste Pfad wird mit Hilfe des im Pathway Hunter Tool implementierten „Shortest Path"- Algorithmus bestimmt (Jungnickel, 2002). Beim Shortest Path-Algorithmus handelt es sich um einen in der Graphentheorie gängigen Algorithmus zur Bestimmung von Abständen innerhalb unterschiedlichster Netzwerke. Der kürzeste Pfad zweier Metaboliten zueinander innerhalb eines metabolischen Netzwerkes beschreibt den kürzesten Reaktionsabstand zweier Metaboliten zueinander. Er besagt, wie viele Einzelschritte notwenig sind, um die beiden Metaboliten miteinander zu verbinden, oder - biochemisch ausgedrückt - ineinander zu überführen. Besitzt der kürzeste Pfad die Länge 1, so sind beide Metaboliten, wie bereits erwähnt, nur durch einen einzelnen enzymkatalysierten Reaktionsschritt voneinander entfernt. Nachfolgend ist beispielhaft ein metabolischer Pfad mitsamt seiner chemischen Strukturformeln angegeben. Er betrachtet die Verbindung zweier Metaboliten aus der Glykolyse und dem Pentose-Phosphat-Weg und kann auch (teilweise) in der vorangegangenen Abbildung 5.2 betrachtet werden.

5.2.3.1.1 Beispielpfad 1 Ausgehend von beta-D-Fructose 1,6-bisphosphate (in grün umrandet) via beta-D-Fructose 6-Phosphate (orange umrandet) zu D-Xylulose 5-Phosphate (blau umrandet). Der Pfad besteht aus zwei Reaktionsschritten (Pfadlänge = 2).

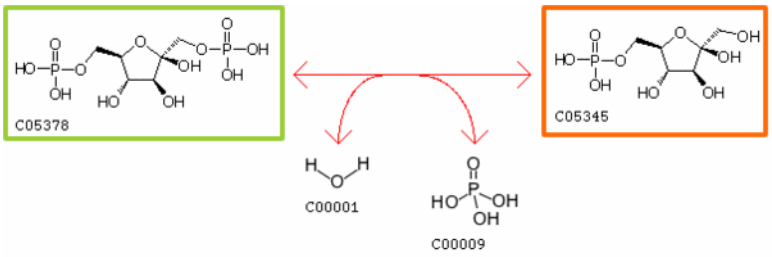

Abbildung 5.3: Beispielpfad 1, Reaktionsschritt 1 (R04780), Quelle: KEGG

Im ersten Reaktionsschritt, wird beta-D-Fructose 1,6-bisphosphate unter Verwendung von Wasser reversibel zu beta-D-Fructose 6-Phosphate und Orthophosphate umgesetzt. Katalysierendes Enzym ist in diesem Falle Hexose Diphosphatase (EC 3.1.3.11). Die Molekülähnlichkeit der betrachteten Metaboliten (in grün und orange gekennzeichnet) ist aufgrund der Abspaltung der Phosphatgruppe sehr hoch.

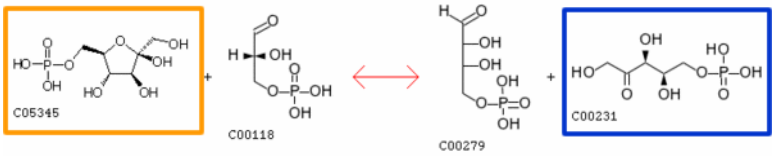

Abbildung 5.4: Beispielpfad 1, Reaktionsschritt 2 (R01830), Quelle: KEGG

Im zweiten Reaktionsschritt wird beta-D-Fructose 6-phosphate unter Verwendung von D-Glyceraldehyde 3-phosphate (C00118) reversibel zu D-Erythrose 4-phosphate (C00279) und dem Endprodukt des betrachteten metabolischen Pfades, D-Xylulose 5-phosphate (blau umrandet) umgewandelt. Katalysiert wird diese Reaktion durch das Enzym Transketolase (EC 2.2.1.1). Es zeigt sich, das die molekulare Ähnlichkeit von beta-D-Fructose 6-phosphate zu D-Xylulose 5-phosphate deutlich geringer ist, als im vorangegangenen Reaktionsschritt.

5.2.3.2 Anzahl kürzester Pfade

Die Anzahl kürzester Pfade ist eine weitere Größe, die der „Shortest Path“- Algorithmus liefert. Sie gibt an, wie viele gleich „kurze“ Pfade zwischen zwei Meta-

boliten innerhalb eines metabolischen Netzwerkes existieren. Dieses ist von biologischer Relevanz, da unter Umständen alternative Reaktionsschritte über unterschiedliche Reaktionspartner führen oder innerhalb der Reaktionen unterschiedliche Substrate Verwendung finden. Die Anzahl der kürzesten Pfade zwischen zwei Metaboliten beschreibt folglich, ob und in welchem Maße eine gleich kurze Alternative für die Umwandlung zur Verfügung steht. Es ist anzunehmen, dass Metaboliten zwischen denen viele alternative Pfade existieren, sozusagen mehrfach gegen Veränderungen und Einflüsse abgesichert sind, als solche Paarungen zwischen denen nur ein einzelner Pfad existiert und Änderungen in der Verfügbarkeit des ersten Metaboliten die Konzentration des zweiten Metaboliten direkt beeinflussen. Die Verfügbarkeit mehrerer alternativer Reaktionswege kann ein Zeichen dafür sein, dass die entsprechende Reaktion für den Organismus von besonderer Bedeutung ist und dass diese selbst bei Ausfall von einzelner Reaktionswege (sei es durch äußere Einflüsse oder durch genetische Mutationen) immer noch in redundanter Form durchgeführt werden kann. Dies kann als erstes Indiz für die Robustheit biologischer Systeme (Kitano, 2004) angesehen werden. Nachfolgendes Beispiel stellt einen solche Pfadalternative für das vorherige Pfadbeispiel von beta-D-Fructose 1,6-bisphosphate zu zu D-Xylulose 5-Phosphate dar.

5.2.3.2.1 Beispielpfad 2 Anstelle über beta-D-Fructose 6-Phosphate verläuft der Pfad über D-Glyceraldehyde 3-phosphate (orange umrandet), einem Metaboliten der uns auch schon im vorangegangen Beispiel als Substrat begegnet ist zu D-Xylulose 5-phosphate (blau umrandet). Die Pfadlänge ist ebenfalls 2 Schritte lang.

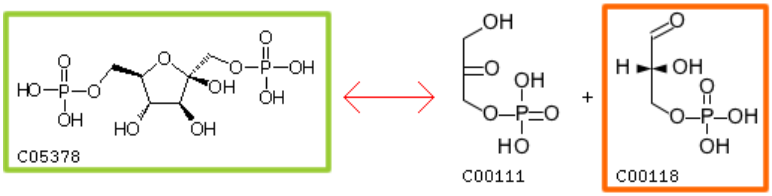

Abbildung 5.5: Beispielpfad 2, Reaktionsschritt 1 (R01070), Quelle: KEGG

Im ersten Reaktionsschritt wird beta-D Fructose 1,6-bisphosphate reversibel zu Glycerone phosphate (C0111) und D-Glyceraldehyde 3-phosphate umgewan-

delt (orange umrandet). Katalysierendes Enzym ist die Fructose-Bisphosphate Aldolase (EC 4.1.2.13). Es fällt auf, dass keine hohe Molekülähnlichkeit zwischen den beiden betrachtenden Metaboliten (grün und orange dargestellt) existiert. Im zweiten Reaktionsschritt wird D-Glyceraldehyde 3-phosphate (orange umrandet) zusammen mit D-Fructose 6-phosphate (C00085) irreversibel zu D-Erythrose 4-phosphate (C00279) und D-Xylulose 5-phosphate (blau umrandet) umgewandelt. Katalysiert wird diese Reaktion, wie im Reaktionsschritt 2 des ersten Beispielpfades auch, durch das Enzym Transketolase (EC 2.2.1.1). Die Transketolase ist, wie im späteren Verlauf dieser Arbeit noch detailliert beschrieben wird, ein wichtiges Bindeglied zwischen der Gykolyse und dem Pentose-Phosphat-Weg, wobei gleich mehrere unterschiedliche Metaboliten ineinander umgesetzt werden können.

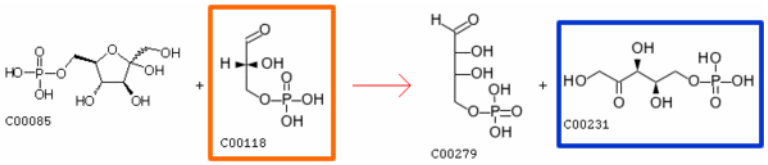

Abbildung 5.6: Beispielpfad 2, Reaktionsschritt 2 (R01067), Quelle: KEGG

Es zeigt sich, dass alternative und gleich kurze metabolische Pfade ihre Wege über unterschiedlichste Reaktionspartner nehmen können, die verschiedene molekulare Ähnlichkeiten besitzen können. Ferner ist festzuhalten, dass die Richtungsabhängigkeit von Einzelreaktionen innerhalb alternativer Pfade nicht zwangsläufig gleichartig ausgeprägt sein muss.

5.2.3.3 Anteil der Pfadlänge am Zitratzyklus

Zyklisch miteinander verbundene Metaboliten stellen einen Sonderfall innerhalb metabolischer Netzwerke dar. Sie sind häufig im Zentralstoffwechsel angesiedelt und sind meistens besonders frequentierte Bestandteile des metabolischen Netzwerkes, die sich durch Selbstregulation beziehungsweise konservatives und robustes Verhalten auszeichnen. Der bekannteste Zyklus in metabolischen Netzwerken ist der Zitratzyklus. Um die Besonderheit der zyklischen Verknüpfung von Metaboliten zueinander zu berücksichtigen, wurde ein Deskriptor entwickelt, der überprüft ob ein gefundener Pfad einen Abschnitt seines Weges über den Zitrat-

zyklus nimmt. Ist dies der Fall, wird die Pfadlänge innerhalb des Zitratzyklus zur Gesamtlänge des gefundenen Pfades in Beziehung gesetzt.

5.2.4 Indirekte theoretische Deskriptoren (auf Einzelschritten berechnet)

Die indirekten Deskriptoren werden mit Hilfe von Detailinformationen aus den jeweiligen einzelnen Reaktionsschritten extrahiert. Anschließend werden die Deskriptoren der Einzelschritte mit Hilfe geeigneter mathematischer Operationen, wie der Berechnung der Extrema (Minimum und Maximum) oder Schwankungsmaßen (Standardabweichung) zusammengefasst. Die indirekten theoretischen Deskriptoren sind in den folgenden Unterkapiteln beschrieben.

5.2.4.1 Anzahl individueller Reaktionen pro Schritt

Eine Umsetzung von einem Metabolit zum anderen kann, wie bereits anhand von Beispielen gesehen, durch unterschiedliche Reaktionen und beteiligte Reaktionspartner ermöglicht werden. Dieser Deskriptor zählt für jeden Einzelschritt, wie viele individuelle Reaktionen für die Umwandlung existieren. Die eindeutige Zuordnung wird hierbei über die Reaktionsnummer, die in KEGG hinterlegt ist, vorgenommen. Dieser Deskriptor ist - ähnlich wie die Anzahl kürzester Pfade - interessant, da mehrere individuelle Reaktionen, welche einen einzigen Umwandlungsschritt katalysieren, möglicherweise als Zeichen für Robustheit und Redundanz gewertet werden können. In anderen Worten: existiert zwischen zwei Metaboliten eine große Anzahl von Reaktionen, die eine Umsetzung ermöglichen, so ist es wahrscheinlicher, dass dieser Weg auch dann gegangen werden kann, wenn unter Umständen ein Substrat nicht zur Verfügung steht. Folgendes Beispiel soll diesen Deskriptor illustrieren. Angeführt sind diejenigen Reaktionswege in *C. glutamicum*, welche die Umwandlung von 2-Oxoglutarate (C00026), einem entscheidenden Metaboliten des Zitratzyklus zu L-Glutamate (C00025), einer Aminosäure in einem einzigen Reaktionsschritt vollziehen. Es zeigt sich, dass die Umsetzung im Stoffwechsel von *C. glutamicum* auf drei Arten stattfinden kann, welche durch individuelle Reaktionsnummern gekennzeichnet sind (R00114, R00248 sowie R00355). Drei verschiedene Enzyme (Glutamate Synthase, Glutamate Dehydrogenase sowie Aspartate Transaminase) können die Umsetzung katalysieren. Im Falle der Glutamate Synthase (Beispiel 1) wird die Aminosäure

L-Glutamate aus 2-Oxoglutarate sowie L-Glutamine und NADPH synthetisiert. Dieser Reaktionsweg ist charakteristisch für einzellige Organismen und verläuft in Reaktionsrichtung von L-Glutamate in irreversibler Form. Bei der Glutamate Dehydrogenase (Beispiel 2) wird Ammonium verstoffwechselt. Diese Reaktion wird mit der Reaktionsnummer R00248 kodiert. Die Aspartate Transaminase (Beispiel 3) stellt den letzten Reaktionsweg dar, bei ihr wird 2-Oxoglutarate unter Verwendung von L-Aspartate in L-Glutamate sowie Oxaloacetate verstoffwechselt.

Beispiel individueller Reaktionen, welche in *C. glutamicum* den Reaktionsschritt vom Metaboliten 2-Oxoglutarate (C00026) zu L-Glutamate (C00025) ermöglichen.

1. R00114 (Glutamate Synthase) [EC 1.4.1.13]:

   2-Oxoglutarate + L-Glutamine + NADPH + H^+ => **2 L-Glutamate** + $NADP^+$

2. R00248 (Glutamate Dehydrogenase) [EC 1.4.1.4]:

   2-Oxoglutarate + NH_3 + NADPH + H^+ <=> **L-Glutamate** + $NADP^+$ + H_2O

3. R00355 (Aspartate Transaminase) [EC 2.6.1.1]:

   2-Oxoglutarate + L-Aspartate <=> **L-Glutamate** + Oxaloacetate

Wie man erkennen kann, werden zur Umwandlung von 2-Oxoglutarate in L-Glutamate in jeder Reaktion unterschiedliche Substrate benötigt. Dies kann ein Indikator dafür sein, dass der Organismus unter unterschiedlichen Umweltbedingungen (beispielsweise Verfügbarkeit von Nährstoffen) unterschiedliche Möglichkeiten der Synthese von L-Glutamate besitzt. Ferner wird ersichtlich, dass nicht alle Reaktionen reversibel sind. Auf die Eigenschaft der Reversibilität als beschreibende Größe wird detailliert im nachfolgenden Kapitel eingegangen.

5.2.4.2 Anteil reversibler Reaktionen

Die Richtungsabhängigkeit von Reaktionen ist bei der Betrachtung von Stoffwechselvorgängen von entscheidender Bedeutung. So können beispielsweise Gleichgewichtszustände dadurch eingestellt werden, dass Reaktionen wahlweise in die eine oder andere Richtung ablaufen können. Dieser Deskriptor greift diesen Sachverhalt

auf und ermittelt wie hoch der Anteil reversibler Reaktionen pro Schritt ist. Der verbleibende Anteil besteht aus gerichteten Reaktionen. Betrachtet man obiges Beispiel, so kann man die Richtungsabhängigkeit der Reaktionen an dem Richtungsanzeigern erkennen. Das Zeichen <=> steht für reversibel ablaufende Reaktionen, während das Zeichen => gerichtet ablaufende Reaktionen kennzeichnet. Hier kommen zu zwei Dritteln reversible ablaufende Reaktionen vor, was einem Anteil von 66% beziehungsweise einem Deskriptorenwert von 0,66 entspricht.

5.2.4.3 Anzahl individueller Enzyme

Dieser Deskriptor greift ab, wie viele unterschiedliche Enzyme pro Reaktionsschritt auftreten. Hierbei wird in allen Reaktionen dieses Schrittes nach unterschiedlichen EC-Nummern gesucht. Treten innerhalb einer Einzelreaktionen Multi-Enzymkomplexe auf, so werden die dazu gehörigen Enzyme einzeln gezählt. Dieser Deskriptor besitzt Ähnlichkeit zur Anzahl von Reaktionen pro Einzelschritt, ist aber nicht deckungsgleich. Generell ist jedoch anzunehmen, dass eine hohe Anzahl unterschiedlicher Enzyme innerhalb eines Reaktionsschrittes dafür spricht, dass der betreffende Reaktionsschritt auch dann durchführbar ist, wenn beispielsweise ein bestimmtes Enzym in seiner katalytischen Funktion gehemmt ist. Ein redundantes Vorkommen von Enzymen, die den selben Reaktionsschritt katalysieren, kann insofern als wichtig gewertet werden, da dieser Schritt gewissermaßen mehrfach gegen Veränderungen abgesichert ist. In diesem Zusammenhang muss allerdings beachtet werden, dass das organismenspezifische Wissen über das Vorhandensein von Enzymen eine stetige Weiterentwicklung erfährt. Die Anzahl unterschiedlicher Enzyme pro Reaktionsschritt wäre im oben angeführtem Beispiel gleich 3.

5.2.4.4 Mittlere Anzahl von Enzymen pro Reaktion

Dieser Deskriptor untersucht, ob Multi-Enzymkomplexe für die Reaktionen eines Schrittes vorliegen und berechnet daraus einen Index. Dabei wird bestimmt, wie viele Enzyme im Schnitt pro Reaktion und Schritt vorkommen. Im letztgenannten Beispiel aus Kapitel 5.2.4.1 ist es recht einfach; hier ist der Wert für den Deskriptor gleich 1, da im Mittel nur ein Enzym pro Reaktion vorliegt. Für diesen Deskriptor wird zusätzlich zum Mittelwert das Maximum, als auch das Minimum abgegriffen.

5.2.4.5 Betrachtung der Gibbs-Energie

Die Gibbs-Energie (oder auch Gibbs-Potenzial genannt) ist eine thermodynamische Größe. Sie kann für chemische Reaktionen bestimmt werden, wobei sie in erster Linie von Druck, Temperatur und der Konzentration der beteiligten Reaktionspartner abhängig ist. Sie gibt an, wie viel Energie (in kJ/mol) unter den gegebenen Randbedingungen für eine Reaktion bei vorgegebenen Edukten und Produkten unter Berücksichtigung der Reaktionsrichtung entweder benötigt oder freigesetzt wird (Cypionka, 2005). Ist die Gibbs-Energie negativ, spricht man von so genannten exergonen Reaktionen, die bei den gegebenen Bedingungen spontan und unter Energiefreisetzung ablaufen. Wenn die Gibbs-Energie positive Werte annimmt bedeutet dies, dass die Reaktion in der definierten Richtung nur unter Zuführung von Energie abläuft. Ist die Gibbs-Energie nahe Null, so handelt es sich um einen Gleichgewichtszustand, was für die Betrachtung metabolischer Systeme bedeutet, dass besonders hier eine Regulation durch biologische Interaktion stattfinden kann. Für die in den Reaktionsnetzwerken vorhandenen Einzelreaktionen sind die Gibbs-Energien von Dr. Kai Hartmann im Rahmen seiner Doktorarbeit (Hartmann, 2007) exemplarisch für den pH-Wert von 7 berechnet worden. Zur Beschreibung des Vorgehens finden sich weiterführende Informationen bei Mavrovouniotis (1991) sowie auf der Website des Unternehmens Chemaxon (`http://www.chemaxon.com`). Die berechneten Potenziale können den Reaktionen eindeutig über die KEGG-Nomenklatur zugeordnet werden. Im Beispiel aus Kapitel 5.2.4.1 konnten für die drei angeführten Wege der Umsetzung von 2-Oxoglutarate zu L-Glutamate unterschiedliche Gibbs-Energien bestimmt werden (Hartmann, 2007). Für das Reaktionsbeispiel 1 (R00114) konnte ein Gibbs-Potenzial von -75,7 kJ/Mol, für das Beispiel 2 (R00248) ein Potenzial von -46,5 kJ/Mol und für das dritte Beispiel (R00355) ein Gibbs-Potenzial von Null bestimmt werden. Es muss an dieser Stelle deutlich darauf hingewiesen werden, dass die berechneten Werte unter theoretisch vordefinierten Bedingungen berechnet worden sind. Über die tatsächlichen in der Zelle vorhandenen Umstände ist nichts bekannt, die ermittelten Gibbs-Potenziale stellen folglich eine erste Näherung an die tatsächlichen Gegebenheiten dar. Deutlich wird allerdings, dass selbst gleich lange metabolische Pfade unterschiedliche energetische Charakteristiken aufweisen können. Es kann folglich vermutet werden, dass in biologischen Systemen unter optimalen Bedingungen bevorzugt solche Reaktionen ablaufen,

welche energetisch eher günstig zu bewerten sind. Ist dies zum Beispiel wegen veränderter Umweltbedingungen (oder infolge von genetischen Mutationen etc.) nicht möglich, so ist der Organismus möglicherweise dazu gezwungen, energetisch kostenintensivere Wege zu gehen, um das Ablaufen bestimmter Reaktionen und - daraus resultierend - unter Umständen sein Überleben sicherzustellen. Die Aktivierung bioenergetisch kostenintensiverer Pfade bei Unterbindung bevorzugter Stoffwechselwege konnte von Rahman und Schomburg im Jahre 2006 für eine vergleichende theoretische Untersuchung von *Bacillus subtilis* und *Bacillus anthracis* beschrieben werden.

Nun jedoch zurück zur Gibbs-Energie als verwendetem Deskriptor. Um zusätzliche Informationen zu erhalten, wurden die Gibbs-Potenziale ferner hinsichtlich ihres Maximums, ihres Minimums, ihres Mittelwertes sowie ihrer Standardabweichung für alle Reaktionen eines betrachteten Umwandlungsschrittes abgeleitet.

5.2.4.6 Metabolitverknüpfung

Metaboliten sind innerhalb ihrer theoretischen Netzwerke, wie im späteren Kapitel 7.2.1 grundlegend beschrieben, nicht gleichartig stark konnektiert. Diese topologische Eigenschaft wurden bereits in anderen Studien untersucht. Bei Metaboliten, welche eine sehr hohen Verknüpfungsgrad aufweisen, handelt es sich häufig um wichtige Metaboliten des Zentralstoffwechsels (Ma und Zeng, 2003a). In solchen Fällen spricht man auch von „Metabolit-Hubs“-, also ausgeprägten Knotenpunkten innerhalb des Netzwerkes, die Bestandteil zahlreicher metabolischer Pfade sind. Im Gegensatz hierzu, existieren auch Metaboliten, welche nur sehr wenige Verknüpfungen zu anderen Metaboliten besitzen. Diese finden sich häufig an der Peripherie des Netzwerkes. Der Verknüpfungsgrad eines Metaboliten ist deshalb eine wichtige Größe, da er für jeden betrachteten Metaboliten gewissermaßen eine Charakterisierung seines Umfeldes liefert. Die Verknüpfungsinformation wurden als Deskriptor zum einen aus den Modellierungen mit dem Pathway Hunter Tool abgeleitet, als auch in generalisierter Form aus der KEGG-Datenbank extrahiert.

5.2.4.7 Anzahl von Pfaden durch den Metaboliten

Ebenfalls wurde durch die Analyse mit dem Pathway Hunter Tool bestimmt, wie viele unterschiedliche metabolische Pfade ihren Weg über einen definierten Metaboliten nehmen, wenn man das gesamte theoretische metabolische Netzwerk unter

den gegebenen Einstellungen der Modellierung betrachtet. Dies bedeutet, man erhält gewissermaßen einen Eindruck darüber, wie frequentiert der betrachtete Metabolit bei einer ganzheitlichen Untersuchung des zugrundeliegenden metabolischen Netzwerkes ist. Anzumerken bleibt, dass dieser Deskriptor eine gewisse konzeptionelle Ähnlichkeit zur Metabolitverknüpfung besitzt.

5.2.4.8 Metabolitladung

Die Ladung eines Knotenpunktes (in unserem Falle eines Metaboliten) ist eine theoretische Größe, welche sich aus dem Verknüpfungsgrad des Metaboliten und der oben genannten Anzahl der Pfade, die ihren Weg über ihn nehmen, zusammensetzt. Die ermittelte individuelle Anzahl von Pfaden wird hierbei durch den Verknüpfungsgrad des Metaboliten dividiert. Anschließend wird dieser Wert durch Division in Beziehung zur durchschnittlichen Ladung des gesamten Netzwerkes gesetzt (siehe Formel 5.9). Hieraus ergibt sich gewissermaßen eine standardisierte Betrachtungsweise. Die Ladung eines Metaboliten innerhalb eines gegebenen Netzwerkes kann wie folgt bestimmt werden.

$$d_{load} = \ln\left(\left(\frac{nPfad_x(in/out)}{nLinks_x(in/out)}\right) / \left(\frac{\sum_{x=i}^{n} nPfad_i(in/out)}{\sum_{x=i}^{n} nLinks_i(in/out}\right)\right) \tag{5.9}$$

Aufgrund der Richtungsabhängigkeit biochemischer Reaktionen kann die Ladung für einen Metaboliten zweifach bestimmt werden. Zum einen für alle Verknüpfungen, welche in ihn hineingehen und zum anderen für alle Verknüpfungen die aus ihm herausgehen. In ihrer Arbeit stellen Rahman und Schomburg im Jahre 2006 die Vermutung auf, dass die Metabolitladung ein entscheidender Hinweis auf die Wichtigkeit eines Metaboliten bei der Betrachtung von metabolischen Netzwerken sein kann. Aus diesem Grund, und um eine vergleichende Untersuchung mit der Metabolitverknüpfung in der integrativen Analyse durchzuführen, wurde die Metabolitladung als letzter theoretischer Deskriptor aus den Modellierungen abgeleitet.

Sämtliche in Kapitel 5.2 genannten theoretischen Deskriptoren wurden aus den im Jahre 2007 durchgeführten Netzwerkmodellierungen abgeleitet und stellen folglich den zu diesem Zeitpunkt aktuellen Wissensstand dar.

5.3 Fusionierung experimenteller und theoretischer Daten

Nach Abschluss der Extraktion von Deskriptoren aus den Rohdaten lagen auf der experimentellen Seite aufgrund der Tatsache, dass die Berechnung sämtlicher Deskriptoren zusätzlich auf der ersten Ableitung der Daten durchgeführt wurde, insgesamt 14 individuelle Deskriptoren vor. Auf der theoretischen Seite finden sich, aufgrund der angesprochenen Verrechnung der indirekten Deskriptoren auf Pfadlänge im Ganzen 144 individuelle beschreibende Größen. Die Deskriptoren sind zusammenfassend in der Tabelle 5.4 aufgeführt. Um eine integrative Analyse experimenteller und theoretischer Datensätze zu ermöglichen, wurde eine einheitliche Datenstruktur entwickelt (siehe Kapitel 5.3.3).

5.3.1 Namenskonvention

Wie bereits angerissen, ist eine - die Metabolitbezeichnung betreffende - allgemein gültige Nomenklatur nicht vorhanden. Innerhalb einzelner Sprachen existieren oft für identische Substanzen zahlreiche Bezeichnungen; von den vorhandenen Schwierigkeiten über Sprachbarrieren hinweg, ganz zu schweigen. Da eine einheitliche Bezeichnung für die Zusammenführung der Metaboliten in der integrativen Analyse unabdingbar ist, wurde ein Standard eingeführt. Als Referenz diente die KEGG-Nomenklatur, die für jede chemische Verbindung einen universalen Bezeichner, die sogenannte „Compound“-Nummer bereithält. Hierzu mussten vor allem die experimentell erfassten Daten hinsichtlich ihrer Bezeichnung in das entsprechende KEGG-Format übersetzt werden, was unter Zuhilfenahme automatisierter Skripte erfolgte.

5.3.2 Auswahl konkurrierender Pfadrichtungen

Nach Berücksichtigung der Namenskonvention wurde in einem weiteren Schritt für alle Paarungen experimentell erfasster Metaboliten untersucht, ob aus der Netzwerkmodellierung Information über einen dazwischen liegenden Pfad vorliegt. Anders ausgedrückt wurde überprüft, welche experimentell bestimmten Metaboliten durch einen theoretischen metabolischen Pfad miteinander verbunden sind. Da die auf den experimentellen Daten berechneten paarweisen Deskriptoren

Tabelle 5.4: Tabellarische Übersicht der aus experimentellen und theoretischen abgeleiteten Deskriptoren.

Experimentelle Deskriptoren	
Pearson'scher Korrelationskoeffizient	
Spearman'scher Rangkorrelationskoeffizient	
Winkelähnlichkeit	
Gleichläufigkeit	
Log10-Verhältnisse	
Sensitivität	
Mutual Information	
Theoretische Deskriptoren	
Kürzester Pfad	
Anzahl kürzester Pfade	
Anteil der Pfadlänge am Zitratzyklus	
Prozentsatz ungerichteter Reaktionen	Verrechnung auf Pfadebene via Max, Min, Mittelwert, Median, Summe und Standardabweichung
Anzahl individueller Enzyme	
Mittlere Anzahl von Enzymen pro Reaktion	
Anzahl unterschiedlicher Gibbs-Energien	
Maximale Gibbs-Energie	
Minimale Gibbs-Energie	
Mittlere Gibbs-Energie	
Standardabweichung der Gibbs-Energien	
Metabolitverknüpfung (ein- und ausgehend)	
Anzahl von Pfaden durch den Metaboliten	
Metabolitladung	

keine Richtungsabhängigkeit aufweisen, musste darauf geachtet werden, dass bei der theoretischen Entsprechung - sofern Hin- und Rückreaktionen gleichermaßen existieren - nur eine Betrachtungsweise beibehalten wird. Zur Erläuterung sei an dieser Stelle ein Beispiel gegeben: betrachten wir zwei hypothetische Zeitreihen der Konzentration zweier Metaboliten A und B. Ihre Korrelation, ihre Winkelähnlichkeit, ihre Gleichläufigkeit als auch ihre Mutual Information sind jeweils gleich, unabhängig davon in welcher Kombination die Metaboliten betrachtet werden. Bei den theoretischen Deskriptoren muss dies allerdings nicht der Fall sein. Vor allem wegen des Vorhandenseins irreversibler Reaktionen in den Reaktionsnetzwerken kommt es zu Besonderheiten, welche separat berücksichtigt werden müssen. So kann beispielsweise der metabolische Pfad von A nach B zwei Reaktionsschritte lang sein, während der umgekehrte Weg 5 Schritte in Anspruch nimmt. Da jedoch für die weitere Datenanalyse eine eindeutige Entsprechung notwendig ist, wurden nachfolgende Kriterien zur Entscheidungsfindung in solchen Spezialfällen entwickelt.

- In der integrativen Analyse findet primär die exponentielle Phase des bakteriellen Wachstums, in der der Aufbau von Biomasse dominiert, für die Berechnung der experimentellen Deskriptoren Anwendung. Konkurrieren zwei Pfadrichtungen miteinander, wurde diejenige verworfen, die von Endprodukten des Stoffwechsels zu Ausgangsprodukten führt. Da reversible Pfade von Endprodukten zu Ausgangssubstraten sehr viel seltener existieren und dann ohnehin meistens länger sind, fand diese Regel nur in sehr seltenen Fällen Anwendung.

- Betreffen die metabolischen Pfade weder Ausgangssubstrate noch Endprodukte, so wurde bei ungleich langen Hin- und Rückreaktionen, die kürzere der beiden Varianten ausgewählt. Dieser Überlegung liegt zugrunde, dass im Allgemeinen kürzere Pfade für die Umwandlung zweier Metaboliten bevorzugt werden.

- Bei gleich langen konkurrierenden Pfadrichtungen wurde diejenige ausgewählt, welche am ehesten in Richtung des Aufbaus von Biomasse verläuft. Ist diese Entscheidung nicht eindeutig zu treffen, wurde diejenige Kombinationen mit dem energetisch günstigeren (entlang des Pfades aufsummierten) Gibbs-Potenzial ausgewählt.

5.3.3 Datenspeicherung

Wie bereits erwähnt, wurde eine einheitliche Datenstruktur entwickelt, die im weiteren Verlauf der Arbeit „Metpair“-Datenstruktur genannt wird. In diesem Datenformat sind sowohl die Rohdaten als auch alle getesteten Vorverarbeitungsschritte, sowie ferner die daraus abgeleiteten Deskriptorensets integriert. Damit ist zu jedem Zeitpunkt eine Reproduzierbarkeit sichergestellt. Zusätzlich ist beschreibende Meta-Information hinterlegt, welche es erlaubt, die Datenstruktur beispielsweise nach Schlüsselworten abzufragen und somit für die mathematisch-statistische Analyse passende Ausgangsdatensätze zu generieren. Ein Beispiel einer solchen Abfrage könnte wie folgt formuliert sein: „Finde diejenigen stark verknüpften Metabolitenpaare aller Fermentationen, welche sich hinsichtlich ihres temporalen Konzentrationsverlaufes möglichst schwankungsarm verhalten“. „Oder finde jene benachbarten Metabolitenpaarungen, welche sich hinsichtlich ihrer paarweisen Prozessähnlichkeit zwischen den Fermentationsexperimenten möglichst stark unterscheiden“. Wie hieraus ersichtlich wird, sind durch die Integration der Meta-Information die Möglichkeiten der Datenabfrage und damit auch der nachfolgenden integrativen Analyse äußerst vielfältig und somit auch auf Fragestellungen, welche sich ausserhalb des Betrachtungssystems dieser Arbeit befinden, anwendbar. Die „Metpair“-Datenstruktur ist innerhalb der MATLAB-Umgebung deklariert, was einen einfachen Datenaustausch, Kompatibilität zu anderen Systemen sowie die Anwendung mathematisch-statistischer Verfahren, wie im nachfolgenden Kapitel 6 beschrieben, erlaubt.

6 Datenanalyse

Während sich das vorangegangene Kapitel 5 damit beschäftigte, wie aus den zum Teil komplexen Rohdaten geeignete Informationen abgeleitet werden konnten, beschäftigt sich dieses Kapitel nun mit der Analyse der gewonnenen Datensätze.

Wie im Vorfeld beschrieben, liegen „experimentelle“ und „theoretische“ Daten zur Analyse vor. Die Datensätze wurden zunächst einzeln, anschließend gemeinsam in Form einer integrativen Datenanalyse untersucht. Dies ist zulässig, da der gleiche Untersuchungsgegenstand (der Metabolit mit seinen individuellen und paarweisen Eigenschaften) sowohl experimentell als auch theoretisch charakterisiert wird. Demzufolge lassen sich die auf beiden Datensätzen abgeleiteten Deskriptoren generell in zwei Gruppen einteilen: erstere die den Metaboliten als solchen charakterisieren und letztere, die einen paarweisen Vergleich zwischen Metaboliten durchführen. Um sicherzustellen, dass die Daten in einem für die weitere Analyse geeigneten Format vorliegen, wurde wie in Kapitel 5.3.3 beschrieben, ein einheitliches Format der Datenspeicherung eingeführt.

Das übergeordnete Ziel der Analyse besteht darin, auffällige Muster innerhalb und zwischen den „experimentellen“ und „theoretischen“ Daten aufzudecken. Oder vereinfacht ausgedrückt: auffällige Zusammenhänge zu finden, die beispielsweise erklären, warum sich ein Metabolit in seinem zeitlichen Konzentrationsverhalten so ausprägt, wie er es tut. Aus diesem Grunde und aufgrund der vergleichsweise niedrigen Stichprobenzahl wurde bei der Datenanalyse auf überwachte Lernverfahren verzichtet, stattdessen fanden Verfahren der unüberwachten, explorativen, Datenstrukturanalyse Anwendung (Hastie et al., 2001). Darüber hinaus wurde nach Zusammenhängen auch mit Hilfe von Korrelationsuntersuchungen, sowie mit visueller Unterstützung durch Scatterplots und anderer Darstellungsformen, gefahndet. Die nachfolgenden Unterkapitel geben in kurzer Form Aufschluss darüber, welche Verfahren hierbei auf den Datensätzen zur Klärung von Auffälligkeiten und Zusammenhängen verwendet wurden.

6.1 Unüberwachte Lernverfahren

Unüberwachte Lernverfahren dienen dazu, einen Datensatz auf Strukturen zu untersuchen und diese zu beschreiben. Im Mittelpunkt der Betrachtungsweise stehen die Messwerte als solche. Zusatzinformationen wie Gruppierungsvariablen werden nicht berücksichtigt. Die Anwendungsmöglichkeiten für unüberwachte Verfahren sind sehr vielfältig, dazu gehören unter anderem: die Zusammenfassung von Objekten in logische Gruppen, das Überprüfen von Gruppenaufteilungen für anschließend durchgeführte überwachte Lernverfahren, die Untersuchung der Heterogenität eines Datensatzes, das Aufspüren von Ausreißern und Fehlern, die Reduktion der Dimension der Eingabedaten und vieles mehr. So unterschiedlich die Anwendungsbereiche unüberwachter Lernverfahren sind, so unterschiedlich sind auch deren mathematische Grundlagen und Konzepte.

Es ist sinnvoll, unüberwachte Lernverfahren zu Beginn eines Datenauswerteprojektes anzuwenden, da beispielsweise wertvolle Informationen über die Konsistenz der Messbedingungen gewonnen werden können. Liegt beschreibende Meta-Information zu den Messwerten vor, die eine Erklärung für gefundene Unregelmäßigkeiten liefert, können diese gegebenenfalls korrigiert und für die weitere Analyse verwendet werden. Kein unüberwachtes Lernverfahren ist universell auf alle Fragestellungen gleichermaßen gut anwendbar. Durch die Vielgestaltigkeit der Verfahren kann es vorkommen, dass unterschiedliche Verfahren auf einem identischen Datensatz unterschiedliche Strukturen aufdecken. Es ist daher unerlässlich, mehrere Verfahren zur Interpretation der Datenstruktur heranzuziehen und miteinander zu vergleichen. Insbesondere wenn - wie in diesem Fall - unterschiedliche Strategien der Datenvorverarbeitung auf den Daten getestet worden sind, ist es sinnvoll die unüberwachte Datenstrukturanalyse iterativ durchzuführen.

6.1.1 Clusteranalyse (CA)

Die Clusteranalyse dient dazu, einzelne Objekte in größere homogene Gruppen (Cluster) zu ordnen und damit begreifbarer zu machen. Ziel des Ansatzes ist es, eine möglichst große Homogenität innerhalb eines Clusters und gleichzeitig eine möglichst große Heterogenität zwischen den Clustern zu erreichen. Es existieren zahlreiche Clusteralgorithmen, welche sich hinsichtlich ihrer Konzeption unterscheiden. Im Rahmen dieser Arbeit wurde schwerpunktmässig der Ward-

Algorithmus (Ward, 1963) verwendet. Bei der Wahl der Distanzmaße wurden verschiedene Ansätze überprüft.

6.1.2 Hauptkomponentenanalyse PCA (Principal Component Analysis)

Projektionsverfahren dienen dazu, hochdimensionale Datensätze in einen Datenraum geringerer Dimensionalität abzubilden. Dieses Vorgehen hat den Vorteil, dass dem Datensatz zugrunde liegende Strukturen im Datenraum niedriger Dimensionalität besser erkannt werden können. Die Hauptkomponentenanalyse (PCA) ist wohl das bekannteste Projektionsverfahren. Sie ist ein multivariates statistisches Verfahren zur Extraktion eines Satzes von unabhängigen und orthogonalen (daher unkorrelierten) Variablen (auch Hauptkomponenten genannt) aus einem höherdimensionalen Datensatz. Die extrahierten Hauptkomponenten sind nach ihrem Anteil an erklärter Varianz geordnet und als Linearkombination aus den zugrunde liegenden Ausgangsvariablen erzeugt worden. Die Zerlegung eines Datensatzes in seine Hauptkomponenten ist daher eine reproduzierbare und reversibel durchführbare Datentransformation. Die Betrachtung eines Datensatzes in seiner Repräsentation durch Hauptkomponenten ermöglicht es, komplexen Sachverhalte besser begreifbar zu machen. Ferner erlaubt sie, den Beitrag der einzelnen Variablen zur den Hauptkomponenten anhand ihrer Ladung abzugreifen, was zusätzliche interpretatorische Möglichkeiten eröffnet.

7 Ergebnisse

Der Ergebnisteil ist hierarchisch aufgebaut und gliedert sich in mehrere aufeinander aufbauende Unterpunkte. Zu Beginn werden zunächst die fehlerkorrigierten und mit unterschiedlichen Verfahren aufbereiteten Daten vor der Ableitung der Deskriptoren einer grundlegenden statistischen Analyse unterzogen, wobei Verfahren der deskriptiven Statistik sowie grundlegende unüberwachte Verfahren der Datenstrukturanalyse (Kapitel 6.1) angewandt wurden. Da die theoretischen Daten das Ergebnis von Modellierungen sind, denen keine Messungen im direkten Sinne zugrunde liegen, beschränkt sich die Untersuchung auf deskriptive statistische Verfahren. Dieser einführenden und grundlegenden Untersuchung schliesst sich die statistische Untersuchung der aus den Daten abgeleiteten Deskriptorensets an. Diese erfolgt bei den experimentellen und theoretischen Deskriptorensets zunächst getrennt, wobei verschiedene Verfahren zum Einsatz kommen. Im Anschluss hieran schließt sich der wohl wichtigste Teil der Analyse an, in dem die Deskriptorensets der experimentellen und theoretischen Daten gemeinsam auf Muster untersucht werden.

7.1 Analyse der experimentellen Ausgangsdaten

An erster Stelle soll eine grundlegende statistische Analyse der experimentellen Ausgangsdaten vor Ableitung der Deskriptorensets gegeben werden.

7.1.1 Betrachtung der Varianzkomponenten

In ihrer Diplomarbeit hat Eliane Frimmersdorf (Frimmersdorf, 2005) bereits eine erste grundlegende statistische Betrachtung auf den Rohdaten durchgeführt. Als wichtigste Ergebnisse sind hierbei die methodische (Tabelle 7.1) und biologische Varianz (Tabelle 7.2) zu nennen. Erstere wurde zu identischen Zeitpunkten durchgeführt, um den Einfluss der Methodik (Zellernte, Zellaufschluss und Derivatisie-

Tabelle 7.1: Methodische Varianz der Fermentationsexperimente

	Exponentielle Phase		Stationäre Phase		
Fermentation	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Gemeinsame Targets
Glucose	7.6%	-	10.7%	-	182
Fructose	13.1%	11.8%	11.5%	12.6%	172
Lactat	11.6%	13.5%	12.2%	11.4%	169
Acetat	13.9%	9.1%	14.8%	12.4%	148
Glutamin	15.8%	13.2%	14.4%	15.6%	137

Tabelle 7.2: Biologische Varianz der Fermentationsexperimente

Fermentationen	Exponentielle Phase	Stationäre Phase	Zeitdifferenz
Acetat (2 Exp.)	35.4%	36.8%	7 Tage
Fructose (2 Exp.)	40.9%	45.8%	30 Tage
Glutamin (2 Exp.)	37.7%	38.1%	7 Tage
Lactat (2 Exp.)	31.5%	26.0%	6 Tage

rung, Quantifizierung) zu evaluieren. Zur Quantifizierung der Abweichung wurde der prozentuale Standardfehler der Replika bestimmt. Bei der Bestimmung der biologischen Varianz wurden die aus den Replika abgeleiteten Mittelwerte beider Experimente zu identischen Zeitpunkten verglichen und daraus der prozentuale Standardfehler bestimmt. Die beiden Tabellen 7.1 und 7.2 sind aus ihrer Diplomarbeit übernommen.

Es zeigt sich, dass die methodische Varianz im Mittel zwischen knapp 11 Prozent in der exponentiellen Phase und knapp 13 Prozent in der stationären Phase anzusiedeln ist. Die jeweiligen Experimente (Exp.1 und Exp.2) unterscheiden sich hinsichtlich ihrer methodischen Varianz nur wenig voneinander, wobei jedoch angemerkt werden muss, dass in der stationären Phase diese Unterschiede geringer sind als in der exponentiellen Phase. Für die Anzucht auf Glucose liegt nur ein Experiment vor, weshalb sich dieser Vergleich erübrigt.

Es wird ferner deutlich, dass die biologische Varianz bedeutend größer ist, als die methodische. Der Vergleich zeigt, dass bei langen Zeiträumen zwischen den betrachteten Wiederholungsexperimenten, die biologische Varianz - am Beispiel der Fructose Fermentationen - besonders hoch ist. Liegen die Fermentationen zeitlich

eng zusammen, so ist die biologische Varianz deutlich geringer, wie am Beispiel der Lactat-Fermentationen. Unter Umständen ist dies auf nicht nachvollziehbare Veränderungen bei der Lagerung der Kulturen zurückzuführen.

7.1.2 Korrelationsanalyse aller Metaboliten inklusive der Unknowns

Je nach Fermentationsexperiment wurde eine unterschiedliche Anzahl von Metaboliten detektiert. Von diesen detektierten Metaboliten sind die meisten auch identifiziert, das heißt über ihre Peaks im Massenspektrum und ihre Retentionszeit eindeutig einer Substanz zugeordnet worden. Es kommt allerdings auch vor, dass keine eindeutige Zuordnung möglich ist. In diesem Falle handelt es sich um sogenannte „Unknowns“, über die zum Zeitpunkt der Analyse keine weiteren Informationen vorhanden waren. Ursächlich kann es sich bei den unidentifizierten Substanzen um noch unbekannte Derivate bereits identifizierter Metaboliten, als auch um gänzlich neue Metaboliten handeln. Um zu untersuchen, ob gewisse Unknows Ähnlichkeiten zu bereits detektierten Metaboliten aufweisen, wurde eine paarweise Korrelationsberechnung auf den (wie in Kapitel 5.1.1.4 beschrieben) nullwert- und ausreißerkorrigierten, mittelwertszentrierten Daten berechnet. Die jeweils 10 ähnlichsten paarweisen Metabolitzeitreihen sind in nachfolgenden Tabellen angegeben. Für die Berechnung der Korrelationskoeffizienten wurde ein Signifikanzniveau von $p < 0{,}05$ angenommen. Das bedeutet, es werden nur ausreichend signifikante Korrelationen berücksichtigt. Die Korrelation wird für diese spezielle Untersuchung bewusst auf der gesamten Zeitreihe berechnet. Ist ein Unknown ein Derivat eines bereits identifizierten Metaboliten, so kann davon ausgegangen werden, dass die beiden Metaboliten sich über die gesamte Zeitreihe sehr ähnlich verhalten. Im Gegensatz hierzu findet bei der späteren integrativen Analyse in Kapitel 7.4, in welcher experimentelle und theoretische Metabolomdaten gemeinsam untersucht werden, eine Einschränkung auf die Phase des exponentiellen Zellwachstums - also einem Abschnitt der Zeitreihe - statt.

7.1.2.1 Acetat-Fermentation

Zur Fütterung auf Acetat gehören zwei Experimente, die im Abstand von 7 Tagen durchgeführt wurden. Folgende Tabelle 7.3 gibt für diese beiden Fermentationen die 10 am stärksten korrelierten Metabolitpaare an. Es zeigt sich, dass bei beiden

Experimenten zwei Unknowns mit den Nummern 51 und 53 am stärksten miteinander korrelieren. Die beiden Unknowns mit den Nummern 109 und A16 stehen ebenfalls in einem starken Zusammenhang in beiden Experimenten. Auffällig ist, dass 2-Phospho D-glycerate (C00631) sowie 3-Phospho D-Glycerate (C00197), beides Metaboliten, die in der Glykolyse bzw. Glukoneogenese eine entscheidende Rolle spielen, jeweils stark mit dem Unknown-67 korrelieren. Hier liegt die Vermutung nahe, dass Unknown-67 ein noch nicht identifiziertes Derivat einer der beiden Metaboliten darstellt. An dieser Stelle sei vorab auch auf Kapitel 7.4.3, verwiesen, in dem detailliert auf Besonderheiten im Stoffwechsel von *C. glutamicum* bei Fütterung mit Acetat eingegangen wird. Die Tatsache, dass sich bereits 3 inhaltliche Übereinstimmungen in beiden Listen beider Experimente finden lassen, ist insofern bedeutsam, da hier aus Gründen der Übersichtlichkeit nur die ersten 10 Kandidaten betrachtet wurden und ferner leichte Abweichungen zu einzelnen Zeitpunkten die Korrelation bereits stark beeinflussen können.

7.1.2.2 Fructose-Fermentation

Die Experimente mit Fructose als Nährmedium wurden zweifach mit einem zeitlichen Abstand von 30 Tagen durchgeführt. Tabelle 7.4 gibt die 10 am stärksten korrelierten Metabolitzeitreihen für beide Experimente an. Es zeigt sich auch hier wieder, dass Unknowns mit identifizierten Metaboliten eine besonderes hohe Korrelation aufweisen. Auffällig ist die Paarung Shikimate (C00493) zu Unknown-70, die zuvor nicht in dieser Deutlichkeit beobachtet werden konnte, nun aber in beiden Experimenten aufzufinden ist. Ferner kann die Paarung bestehend aus den Unknowns A16 und 109 - wie auch schon bei den Acetat-Fermentationen - in beiden Experimenten wiedergefunden werden. Weiterhin ist auffällig, dass unter Fütterungsbedingungen mit Fructose auffallend deutliche Korrelationen zwischen identifizierten Metaboliten gefunden werden können. Hierzu zählen die Paarungen zwischen Glycerol 1-phosphate (C00623) zu Glycerone phosphate (C00111), sowie zwischen beta-D-Fructose (C02336) und Shikimate.

Für die erste Paarung ist als besonders interessant anzusehen, dass der Metabolit Glycerol 1-phosphate - obwohl er in der Fructose-Fermentation messtechnisch nachgewiesen werden konnte - bisher in keiner theoretischen Repräsentation von *C. glutamicum* vorhanden ist. Weder die Existenz des Metaboliten, noch die Existenz eines katalysierenden Enzyms, wurden bisher beschrieben. Dies gilt für die

Tabelle 7.3: Korrelierte Metabolitzeitreihen bei der Anzucht auf Acetat

Acetat-Fermentation 050520			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	Unknown-51	Unknown-53	0,997
2	UnknownA16	Unknown-109	0,996
3	2-Phospho-D-glycerate	Unknown-67	0,987
4	L-Isoleucine	L-Proline	0,961
5	5-Oxoproline	Unknown-76	0,961
6	Unknown-18	Unknown-58	0,955
7	N-Acyl-L-glutamine	Unknown-89	0,951
8	AMP	Unknown-78	0,949
9	L-Tyrosine	D-Phenylalanine	0,943
10	Unknown-24	Unknown-40	0,942

Acetat-Fermentation 050530			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	Unknown-51	Unknown-53	0,996
2	2-Phospho-D-glycerate	Unknown-67	0,986
3	L-Alanine	Unknown-109	0,984
4	Unknown-35	Unknown-58	0,969
5	UnknownA16	Unknown-109	0,968
6	AMP	Unknown-39	0,957
7	3-Phospho-D-glycerate	Unknown-67	0,953
8	L-Threonine	Unknown-18	0,950
9	D-Aspartate	Unknown-39	0,946
10	AMP	Unknown-22	0,943

metabolischen Netzwerke basierend auf der Bielefelder-Annotation (Kalinowski et al., 2003), der Kyowa Hakko- Annotation (Ikeda und Nakagawa, 2003) als auch für die im Rahmen dieser Arbeit erstellten Reaktionsnetzwerke VGL1 und VGL2. Auch in anderen Organismen des gleichen Genus *Corynebacterium* kann kein Hinweis diesbezüglich gefunden werden. Der zweite Metabolit der Paarung, Glycerone phosphate, ist hingegen für *C. glutamicum* vorhanden und in der Theorie beispielsweise über den Metaboliten beta-D-Fructose 1,6-bisphosphate (C05378) enzymatisch mit dem Glykolyse-Stoffwechsel verbunden.

Die hohe Korrelation beider Metaboliten kann unter Umständen ein Zeichen dafür sein, dass beide durch eine bisher nicht in *C. glutamicum* annotierte Reaktion direkt ineinander überführt werden können. Die Datenbank KEGG liefert hierzu mögliche Erklärungsansätze. Glycerol 1-phosphate kann mit Glycerone phosphate durch einen einzigen Reaktionsschritt verbunden sein. Katalysierendes Enzym ist in diesem Fall Glycerol-1-phosphate Dehydrogenase (EC 1.1.1.261), welches unter Verwendung von NADH bzw. NADPH beide Metaboliten reversibel verbindet (Reaktionsnummern R05679 und R05680). Besagtes Enzym konnte in Archaebakterien - wobei an dieser Stelle exemplarisch mit *Aeropyrum pernix* ein Vertreter genannt werden soll - nachgewiesen werden. Eine andere Möglichkeit, wie Glycerol 1-phosphate theoretisch synthetisiert werden kann, ist beispielsweise durch eine Verbindung mit dem Metaboliten Glycerol (C00116). Drei verschiedene Enzyme können eine Reaktion zwischen den beiden Partnern katalysieren: erstes ist das Enzym Diphosphate-glycerol Phosphotransferase (EC 2.7.1.79), welches bisher nur in der Wanderratte (*Rattus norwegicus*) nachgewiesen werden konnte und daher eher unwahrscheinlich ist. Zweites ist das Enzym Glycerol-1-phosphatase (EC 3.1.3.21), welches in Pilzen wie *Saccharomyces cerevisiae* vorkommt. Drittes ist das Enzym Phosphoglycerol Geranylgeranyltransferase (EC 2.5.1.41), welches bisher überhaupt nur in zwei Organismen wie beispielsweise dem methanproduzierenden Bakterium *Methanobacterium thermoautotrophicum* nachgewiesen werden konnte. Zusammengefasst muss gesagt werden, dass aufgrund der hohen Korrelation zwischen Glycerol 1-phosphate und Glycerone phosphate einiges für das Vorhandensein von Glycerol-1-phosphate Dehydrogenase als verbindendes Enzym spricht. Dieses, als auch die anderen theoretisch denkbaren Wege der Synthese von Glycerol 1-phosphate, sollte jedoch in weiteren Studien detaillierter untersucht werden.

In der zweiten hochkorrelierten Paarung sind beta-D-Fructose und Shikima-

te im metabolischen Netzwerk durch einen 6 Schritte langen Pfad miteinander verbunden.

7.1.2.3 Glutamin-Fermentation

Zur Fermentation auf Glutamin gehören zwei Experimente, die im Abstand von 7 Tagen voneinander gemessen worden sind. Bei der Anzucht auf Glutamin als Nährmedium wächst *C. glutamicum* am langsamsten im Vergleich zu den anderen im Rahmen dieser Arbeit betrachteten Fermentationen. Nachfolgende Tabelle 7.5 gibt auch hier die 10 am stärksten korrelierenden Metabolitprofile an. Die Paarungen von Shikimate zu Unknown-70 sowie von 2-Phospho-D-glycerate zu Unknown-67 finden sich in beiden Experimenten als hoch korreliert. Eine hohe Korrelation kann auch zwischen den identifizierten Metaboliten 2-Phospho D-glycerate und 3-Phospho D-glycerate festgestellt werden. Beide sind sind wichtige Bestandteile der Glykolyse bzw. Glukoneogenese und dort nur durch das Enzym Phosphoglycerate Mutase (EC 5.4.2.1) reversibel verbunden. Ebenfalls zwischen den Metaboliten beta-D-Glucose 6-phosphate (C01172) und beta-D-Fructose 6 phosphate (C05345) - in der Glykolyse nur durch das Enzym Glucose-6-phosphate Isomerase (EC 5.3.1.9) reversibel verbunden - existiert eine hohe Korrelation. Über die Ursachen einer hohen Prozessähnlichkeit im Konzentrationsverhalten dieser beiden Metaboliten sei im späteren Kapitel 7.4.3.1 unter Zuhilfenahme von Transkriptomuntersuchungen detailliert eingegangen. Erstaulicherweise zeigen auch Paare von Aminosäuren eine hohe Ähnlichkeit zueinander. So sind beispielsweise die Paarungen L-Homoserine (C00263) und L-Valine (C00183) sowie L-Isoleucine (C00407) und L-Proline (C00148) zueinander sehr hoch korreliert. Auf die möglichen Ursachen einer hohen Prozessähnlichkeit zwischen Aminosäuren wird im Detail in Kapitel 7.4.2.6 eingegangen.

7.1.2.4 Lactat-Fermentation

Auf Lactat als Nährmedium wurden zwei Experimente im Abstand von 6 Tagen durchgeführt. Damit stellen die Lactat-Experimente diejenigen Experimente dar, die im kürzesten zeitlichen Abstand zueinander durchgeführt worden sind und welche die geringste biologische Varianz aufweisen. Auch hier findet sich unter den 10 stärksten Korrelationen eine Paarung zwischen Aminosäuren. Das Paar setzt sich aus L-Lyxose (C01508) und L-Arabinose (C00259) zusammen und ist

Tabelle 7.4: Korrelierte Metabolitzeitreihen bei der Anzucht auf Fructose

Fructose-Fermentation 050602			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	Glycerol 1-phosphate	Glycerone phosphate	0,986
2	UnknownA16	Unknown-109	0,986
3	Shikimate	Unknown-70	0,986
4	AMP	Unknown-8	0,979
5	L-Homoserine	Unknown-49	0,972
6	Unknown-101	Unknown-123	0,972
7	L-Homoserine	Unknown-49	0,972
8	3-Phospho-D-glycerate	Unknown-67	0,972
9	Unknown-78	Unknown-8	0,972
10	Unknown-84	Unknown-97	0,972

Fructose-Fermentation 050809			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	Shikimate	Unknown-70	0,995
2	UnknownA16	Unknown-109	0,989
3	UnknownA16	Unknown-108	0,984
4	Unknown-108	Unknown-109	0,984
5	L-Alanine	Unknown-19	0,984
6	L-Alanine	Unknown-27	0,978
7	2-Oxoglutarate	Unknown-34	0,978
8	beta-D-Fructose	Shikimate	0,973
9	Unknown-69	Unknown-70	0,973
10	Maltose	Unknown-120	0,967

Tabelle 7.5: Korrelierte Metabolitzeitreihen bei der Anzucht auf Glutamin

Glutamin-Fermentation 050627			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	Unknown-56	Unknown-95	1
2	beta-D-Glucose 6-phosphate	UnknownA9	1
3	D-Glucono-1,5-lactone	Unknown-81	0,974
4	Shikimate	Unknown-70	0,969
5	2-Phospho-D-glycerate	Unknown-67	0,952
6	L-Homoserine	L-Valine	0,945
7	2-Phospho-D-glycerate	3-Phospho-D-glycerate	0,942
8	L-Isoleucine	L-Proline	0,935
9	L-Alanine	Unknown-19	0,929
10	(S)-Malate	L-Valine	0,925

Glutamin-Fermentation 050714			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	Shikimate	Unknown-70	0,997
2	beta-D-Glucose 6-phosphate	beta-D-Fructose 6-phosphate	0,971
3	L-Alanine	Unknown-19	0,968
4	2-Phospho-D-glycerate	Unknown-67	0,966
5	2-Phospho-D-glycerate	3-Phospho-D-glycerate	0,965
6	Unknown-19	Unknown-22	0,956
7	2-Phospho-D-glycerate	Unknown-96	0,951
8	L-Alanine	Unknown-27	0,948
9	Phosphoenolpyruvate	Unknown-96	0,945
10	Glutarate	UnknownA4	0,945

gleichermaßen in beiden Experimenten zu finden. Ebenfalls in beiden Experimenten findet sich die signifikant korrelierte Paarung (S)-Lactate (C00186) - dem Ausgangssubstrat - und dem Unknown D-3. Eine weitere auffällig hohe Korrelation besteht zwischen den Metaboliten (S)-Malate (C00149) und 6-Phospho-D-gluconate (C00345) und kann ebenfalls in einem der beiden Fermentationsexperimente gefunden werden.

Die bereits in den Fermentationen von Acetat, Fructose und Glutamin existierende hochkorrelierte Paarung von Shikimate und Unknown-70 kann auch hier - in einem der beiden Experimente - nachgewiesen werden.

7.1.2.5 Glucose-Fermentation

Für die Fermentation auf Glucose existiert nur ein Experiment. Es zeigt sich erneut, dass einige nicht identifizierte Metaboliten hohe Korrelationen zu identifizierten Metaboliten aufweisen, wie zum Beispiel an der Paarung des Metaboliten Sucrose (C00089) zu Unknown-121. Die bereits in anderen Fermentationen gefundene Paarung der Unknowns A16 und 109 findet sich auch hier wieder.

Bei den identifizierten Metaboliten besteht eine starke Korrelation zwischen D-Ribose 5-Phosphate (C00117) und D-Xylulose 5-phosphate (C00231), die im metabolischen Netzwerk von *C. glutamicum*, genauer gesagt im Pentose-Phosphat-Weg, durch das Enzym der Transketolase (EC 2.2.1.1) miteinander verbunden sind. An dieser Stelle sei besonders auf die weiterführenden Untersuchungen der Prozessähnlichkeiten unter Heranziehung von Transkriptomdaten in Kapitel 7.4.3.1 hingewiesen. Diese konnten zeigen, dass der Pentose-Phosphat-Weg unter Fütterungsbedingungen mit Glucose vergleichsweise stark frequentiert ist. Eine hohe Korrelation ist auch zwischen den Zeitreihen der beiden Zuckern D-Mannose (C00159) und beta-D-Glucose (C00221) feststellbar.

7.1.3 Zusammenfassung

Für die integrative Analyse theoretischer und experimenteller Metabolomdaten können die Unknowns wegen ihrer fehlenden Zuordnung zu einer Substanz nicht verwendet werden. Deshalb und aufgrund der Tatsache, dass der Anteil von unidentifizierten an allen gemessenen Metaboliten bis zu 30% beträgt, ist es wichtig, das Verhalten von Unknowns zu bereits bekannten Metaboliten im Vorfeld zu untersuchen. Die Analyse ergab, dass über Fermentationen hinweg wiederkehren-

Tabelle 7.6: Korrelierte Metabolitzeitreihen bei der Anzucht auf Lactat

Lactat-Fermentation 050902			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	L-Arabinose	L-Lyxose	0,995
2	Shikimate	Unknown-70	0,995
3	(S)-Lactate	Unknown-D3	0,995
4	L-Homoserine	Unknown-25	0,989
5	beta-D-Glucose	D-Mannose	0,984
6	D-Glucose	D-Mannose	0,984
7	Homocysteine	Unknown-23	0,978
8	L-Alanine	Unknown-19	0,967
9	Uridine	L-Lysine	0,967
10	UMP	Unknown-89	0,967

Lactat-Fermentation 050906			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	UnknownA17	Unknown-110	0,993
2	Unknown-89	Unknown-91	0,986
3	(S)-Malate	6-Phospho-D-gluconate	0,979
4	Cytosine	Unknown-21	0,979
5	L-Homoserine	Unknown-25	0,979
6	Unknown-84	Unknown-88	0,979
7	Unknown-84	Unknown-91	0,979
8	L-Arabinose	L-Lyxose	0,972
9	UnknownA16	Unknown-109	0,972
10	(S)-Lactate	Unknown-D3	0,965

Tabelle 7.7: Korrelierte Metabolitzeitreihen bei der Anzucht auf Glucose

Glucose-Fermentation 050815			
Nr.	Metabolit A	Metabolit B	Spearman'sche Korrelation
1	Sucrose	Unknown-121	0,996
2	UnknownA16	Unknown-109	0,993
3	D-Mannose	beta-D-Glucose	0,989
4	5-Oxoproline	Unknown-76	0,989
5	L-Arabinose	L-Lyxose	0,986
6	Unknown-84	Unknown-91	0,982
7	D-Ribulose 5-phosphate	Unknown-95	0,982
8	Sucrose	Unknown-D5	0,982
9	D-Ribose 5-phosphate	D-Xylulose 5-phosphate	0,971
10	5-Aminolevulinate	Unknown-67	0,968

de Paarungen hoher Prozessähnlichkeit zwischen identifizierten Metaboliten und Unknowns existieren. So findet sich beispielsweise die Paarung von Shikimate und Unknown-70 mehrfach unter den stärksten Korrelationen. Besonders für jene Unknowns sollte sich eine weiterführende Analyse der zugehörigen Massenspektren und Retentionszeiten aus der GC/MS-Analyse anschließen. Ergibt diese in der Zwischenzeit, dass es sich um Derivate der identifizierten Metaboliten handelt, so muss deren Pseudointensität um den Betrag der jeweiligen Derivate erhöht werden. Kurz vor Fertigstellung dieser Arbeit konnte für Unknown-67, welches mehrfach in hohen Korrelationen mit dem Metaboliten 3-Phospho D-glycerate auftrat, festgestellt werden, dass es sich tatsächlich um ein Derivat desselben handelt (persönliche Kommunikation E. Frimmersdorf).

Gelänge jedoch aufgrund der Korrelationsuntersuchungen die Identifizierung von neuen Metaboliten, so wäre dies sicherlich für das Verständnis der systemischen Zusammenhänge durch das Auffüllen der immer noch zahlreichen Lücken von großer Bedeutung. Bemerkenswert ist weiterhin, dass bereits für einige Paarungen identifizierter Metaboliten deutliche Korrelationen auf Basis der gesamten Zeitreihe festzustellen sind. Dies gilt es in der integrativen Analyse anhand der alleinigen Betrachtung der exponentiellen Wachstumsperiode, in welcher die Produktion von Biomasse das vordringlichste Ziel ist, zu verifizieren.

7.1.4 Datenstrukturanalyse innerhalb der Fermentationexperimente

Die Clusteranalyse wurde in diesem Zusammenhang verwendet, um ähnliche Strukturen sowohl im globalen zeitlichen Verhalten des Stoffwechsels, als auch im zeitlichen Verhalten einzelner Metaboliten herauszuarbeiten. Als Cluster-Algorithmus findet der Ward-Algorithmus Anwendung (Ward, 1963). Grundsätzlich bieten sich zwei Möglichkeiten an, die zeitlich aufgelösten Konzentrationsdaten zu clustern. Zum einen über alle Zeitpunkte hinweg, bei der die Länge der zu clusternden Vektoren der Anzahl von Metaboliten entspricht. Zum anderen die Clusterung in zeitlicher Dimension, wobei die Vektoren den Konzentrationszeitreihen der einzelnen Metaboliten entsprechen.

7.1.4.1 Clusterung der Messzeitpunkte

Die Clusterung über alle Metaboliten hinweg hat das Ziel, temporale Veränderungen im globalen Stoffwechsel von *C. glutamicum* zu detektieren. Die Ergebnisse der Clusterung zeigen, dass zeitliche Abschnitte zusammengefasst werden, welche gut mit den physiologischen Wachstumsphasen übereinstimmen. Eine beispielhafte Clusterung - unter Verwendung medianzentrierter lognormierter Daten der Glucose-Fermentation und dem Spearman'schen Rangkorrelationskoeffizienten als Abstandsmaß - veranschaulicht die Abbildung 7.1. Ähnliches Verhalten ist auf den anderen Fermentationsdaten ebenfalls festzustellen.

Ein großer Cluster umfasst den Zeitabschnitt von 0 bis 360 Minuten nach Beginn der Fermentation. Er kann in zwei Untergruppen unterteilt werden, von denen einer von 0 bis 180 Minuten reicht und somit ungefähr der Lag-Phase zugeordnet werden kann. Der zweite Subcluster reicht von 240 bis 360 Minuten und kann der Übergangsphase zur exponentiellen Wachstumsphase zugeordnet werden. Der zweite große Cluster reicht von 420 Minuten bis zum Ende der Fermentation bei 840 Minuten. Auch er kann auch in zwei Subcluster unterteilt werden, von denen der erste von 720 - 840 Minuten exakt die stationäre Wachstumsphase beschreibt. Der verbleibende Subcluster beschreibt ungefähr die exponentielle Wachstumsphase, wobei im Falle der Glucose-Fermentation unabhängig von diesem Vorgehen unter Heranziehung der optischen Dichte (OD) sowie anderen fermentationsspezifischen Parametern eine Zeitspanne von 480 bis 660 Minuten bestimmt werden konnte. Es zeigt sich also, dass die physiologischen Wachstumsphasen re-

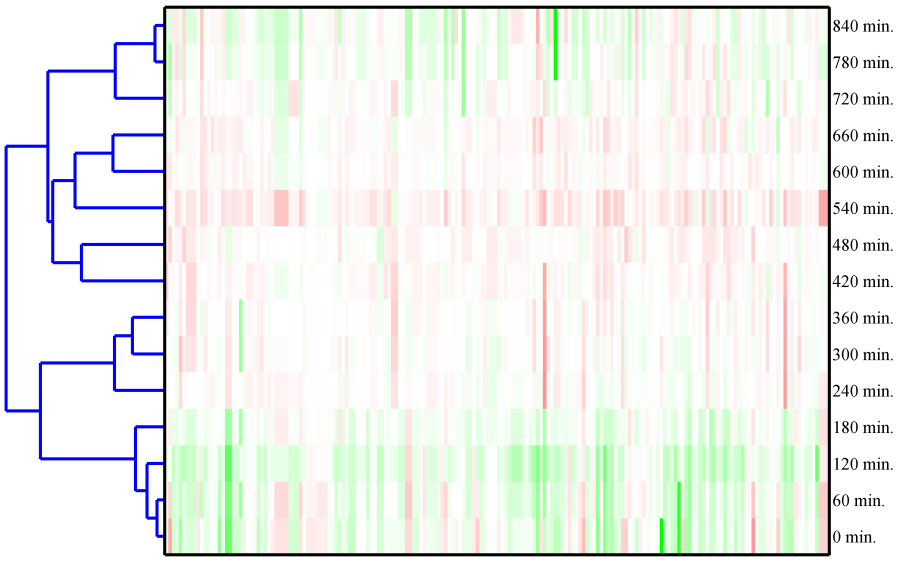

Abbildung 7.1: Clustergramm der Glucose-Fermentation in zeitlicher Dimension

lativ genau aus den zugrunde liegenden Konzentrationsdaten abgeleitet werden können. Diese Ergebnisse zeigen eindrucksvoll, dass sich der gesamte Stoffwechsel von *C. glutamicum* während der Fermentationsexperimente keineswegs gleichartig verhält, sondern grundlegend ändert.

7.1.4.2 Clusterung der Konzentrationszeitreihen

Analog kann auch in Metabolit-Dimension geclustert werden. Das bedeutet, dass die zu clusternden Vektoren durch die Konzentrationszeitreihen der einzelnen Metaboliten repräsentiert werden. Als Vorverarbeitung wurden erneut die medianzentrierten, lognormierten Daten verwendet. Geclustert wurde über die gesamte Länge der Zeitreihe. Aufgrund des großen Umfangs der Clustergrammdarstellungen beschäftigt sich diese Analyse beispielhaft mit jeweils einer Fermentation auf den Ausgangssubstraten Glucose (Experiment 050815), Fructose (Experiment 050602), Acetat (Experiment 050520), Lactat (Experiment 050902) sowie Glutamin (Experiment 050714). Die Darstellungen sind so zu interpretieren, dass die grünlichen Farbtönen relativ geringen Konzentrationen und die roten Farbtöne relativ hohen Konzentrationen entsprechen. In der Clustergrammdarstellung

sind die Metaboliten zeilenweise angeordnet, wobei sie in ihrer Reihenfolge der Clusterzuordnung folgen. Die Clusterzuordnung kann über das am linken Rand angebrachte Dendrogramm nachvollzogen werden. Auf der Abszisse erstrecken sich von links nach rechts die jeweiligen Messzeitpunkte, die aufsteigend in ihrer chronologischen Abfolge dargestellt sind.

7.1.4.2.1 Untersuchung der Glucose-Fermentation

Die Untersuchung des Clustergramms der Glucose-Fermentation (Abbildung 7.2) liefert ein interessantes Bild, denn es können deutlich 3 große Strukturen in den Daten festgestellt werden. Im oberen Teil des Clustergramms erkennt man jene Metabolitprofile, die zu Beginn der Fermentation relativ hohe Werte (rötliche Farbtöne) besitzen, bis sie gegen Ende des Experimentes eher niedrigere Werte annehmen. In dieser Gruppe, die sich gewissermaßen durch kontinuierliches Leerlaufen kennzeichnet, findet sich beispielsweise der Metabolit beta-D Glucose (C00221), welcher das vorhandene Ausgangssubstrat darstellt.

Die restlichen Metaboliten werden von einem großen Cluster zusammengefasst, welcher anhand des Dendrogrammes in zwei annähernd gleich große Gruppen unterteilt werden kann. Der erstere der beiden Subcluster - im Abschnitt von „Unknown 18“ bis zum Metaboliten 5-Aminolevulinate (C00430) - enthält vornehmlich Metabolitprofile, die mit relativ niedrigen Konzentrationen in die Fermentation starten und kontinuierlich bis zum Ende des Experimentes in ihrer Konzentration ansteigen. Zu diesem Cluster stetiger Akkumulation gehört die Mehrheit aller detektierten Aminosäuren wie L-Serine (C00065), L-Isoleucine (C00407), L-Homoserine (C00263), L-Tryptophan (C00078), L-Lysine (C00047), L-Aspartate (C00049), L-Homocysteine (C00155), L-Phenylalanine (C00079) und L-Tyrosine (C00082). Aminosäuren stellen Endprodukte des Stoffwechsels dar und werden - unter anderem - zum Aufbau von Biomasse benötigt. Der zweite, deutlich erkennbare, Subcluster (zu finden im Abschnitt von „Unknown 33“ bis zum Ende des Diagramms) enthält Metaboliten, deren Profile ebenfalls mit relativ niedrigen Konzentrationen beginnen, jedoch aber zu Beginn der stationären Wachstumsphase wieder in ihrer Konzentration abnehmen. In ihrem zeitlichen Verlauf ähneln Metaboliten dieses Clusters dem zeitlichen Verlauf der optischen Dichte (siehe Abbildung 4.4). In diesem Cluster finden sich beispielsweise Pentose-Phosphate wie D-Ribulose 5-phosphate (C00199), D-Ribose 5-phosphate (C00117) und D-Xylulose 5-phosphate (C00231) sowie Metaboliten aus der Glykolyse wie beta-

D-Glucose 6-phosphate (C01172), beta-D-Fructose 6-phosphate (C05345) oder beta-D-Fructose 1,6-bisphosphate (C05378), was ein Hinweis darauf sein könnte, dass die Aktivität der Glykolyse und des Pentose-Phosphat-Weges signifikant mit Eintritt in die stationäre Wachstumsphase zurückgefahren wird. Zusammengefasst kann gesagt werden, dass die Zuordnung eines Metaboliten anhand der Clusteranalyse (besonders am Beispiel der Aminosäuren) bereits eine erste grobe Einschätzung seiner Funktion und Position im metabolischen Netzwerkes erlaubt. Beginn und Ende der exponentiellen Wachstumsphase, welche - wie bereits in den Kapiteln 4.2.1 und 7.1.4.1 beschrieben - ungefähr in einem Bereich von 480 bis 720 Minuten anzusiedeln sind, können in der Clustergrammdarstellung deutlich als Zeitpunkte gravierender Veränderungen (besonders in Cluster 2 und 3) erkannt werden.

7.1.4.2.2 Untersuchung der Fructose-Fermentation

Die exponentielle Wachstumsphase unter Fütterungsbedingungen mit Fructose kann mit Hilfe der Untersuchung der optischen Dichte auf den Bereich von ca. 360 bis 540 Minuten nach Beginn der Fermentation festgelegt werden. In der Clustergrammdarstellung (Abbildung 7.3) ist dieses Zeitintervall deutlich als Phase erhöhter Konzentration vieler Metaboliten zu erkennen. Insgesamt betrachtet sieht die Struktur, verglichen zur Fermentation mit Glucose deutlich heterogener aus. Jene drei - beispielhaft in der Glucose-Fermentation festgestellten Cluster zeitlichen Verhaltens - lassen sich zwar feststellen, wenn aber auch bei weitem weniger deutlich. Hinzu kommt, dass sich einige wenige Metaboliten (beginnend vom obersten Eintrag „Unknown A11“ bis zu beta-D-Glucose 6-phosphate (C01172) dergestalt verhalten, dass sie mit vergleichbar hohen Konzentrationen zu Beginn des Fermentationsexperimentes vorliegen, bis zum Ende der expontentiellen Phase kontinuierlich abnehmen und danach wieder in ihrer Konzentration ansteigen. Vom Metaboliten Oxalate (C00209) bis hin zu L-Alanine (C00041) finden sich jene Metaboliten, die kontinuierlich bis zum Ende der exponentiellen Wachstumsphase ansteigen und anschließend in ihrer Konzentration abnehmen. Ein deutlicher und großer Cluster, welcher in der Darstellung von den unidentifizierten Substanzen „Unknown 99“ bis „Unknown D2“ flankiert wird, enthält alle Metaboliten, die kontinuierlich bis zum Ende des Fermentationsexperimentes in ihrer Konzentration ansteigen. Auch hier bestätigt sich wieder der Sachverhalt, dass mit L-Ornithine (C00077), L-Aspartate (C00049), L-Homocysteine (C00155), L-

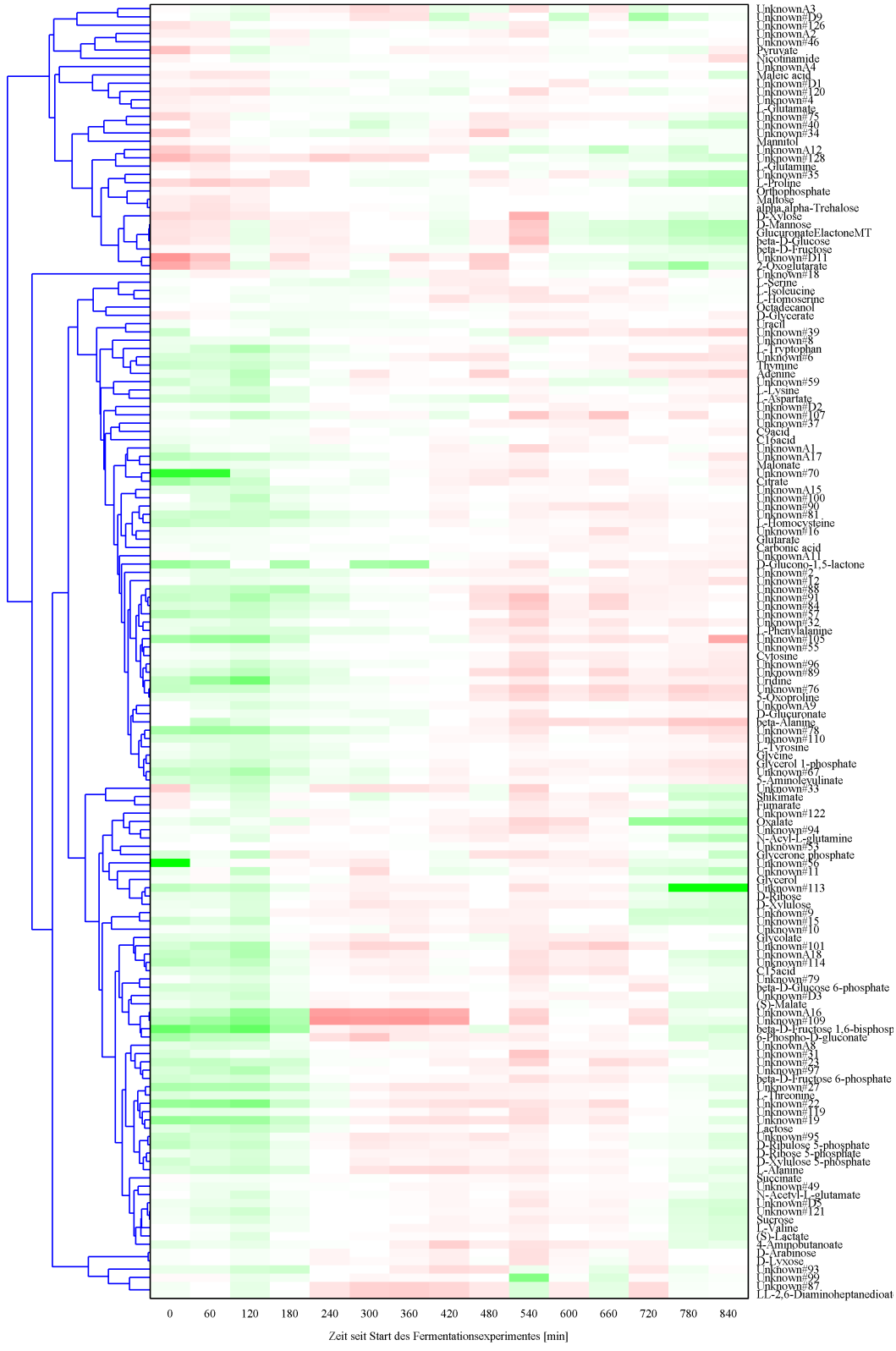

Abbildung 7.2: Clustergramm der Metaboliten aus der Glucose-Fermentation (050815)

Tyrosine (C00082) und L-Lysine (C00047) zahlreiche Aminosäuren diesem Cluster zuzuordnen sind. Ein ebenfalls deutlicher Cluster, welcher jene Metaboliten enthält, die ihre Konzentration beginnend vom Start des Fermentationsexperiments kontinuierlich verringern, findet sich von der Substanz „Unknown A16“ bis hin zum letzten Metaboliten der Liste, beta-D-Fructose 6-Phosphate (C05345). Die hohe Anzahl nicht identifizierter Substanzen in diesem Cluster erschwert allerdings die Interpretation enorm.

7.1.4.2.3 Untersuchung der Acetat-Fermentation

Vergleicht man die Clusterung der Glucose-Fermentation mit der Acetat-Fermentation (050520) in Abbildung 7.4, so fällt auf, dass auch hier eine Einteilung in 3 Gruppen, jedoch weniger deutlich, festgestellt werden kann. Der erste Cluster - oben in der Darstellung bis hin zum Metaboliten 5-Oxoproline (C01879) - enthält schwerpunktmässig Metaboliten, welche mit relativ geringer Konzentration starten und tendenziell bis zum Ende des Experimentes in ihrer Konzentration zunehmen. Beginnend ab der unidentifizierten Substanz „Unknown 5“ bis hin zum Metaboliten 2-Oxoglutarate (C00026) erstreckt sich der zweite große Cluster. Er enthält alle Metaboliten, die zu Beginn des Experimentes mit relativ hoher Konzentration starten, dann entweder kontinuierlich bis zum Ende des Experimentes in ihrer Konzentration abnehmen, oder wie in einem Subcluster - abgegrenzt von „Unknown 5“ bis 2-Phospho-D-Glycerate (C00631) - interessanterweise nach Abschluss der exponentiellen Phase in ihrer Konzentration erneut zunehmen. Dies ist insofern interessant, da sich in dem erwähnten, vergleichsweise kleinen Subluster neben 2-Phospho-D-Glycerate (C00631) auch 3-Phospho-D-Glycerate (C00197) befindet. Beide sind wichtige Metaboliten in der Glykolyse bzw. Glukoneogenese. Ihr Konzentrationsanstieg in der stationären Phase könnte unter anderem durch den Aktivität der Glukoneogenese zu erklären sein. Über die mögliche Aktivität der Glukoneogenese, besonders unter Fütterungsbedingungen mit Acetat, ist bereits in anderen Studien diskutiert worden (Wendisch et al., 2000). Im Rahmen der integrativen Analyse wird in Kapitel 7.4.3 ausführlich auf diesen Sachverhalt eingegangen.

Zuletzt folgt auch in dieser Fermentation ein Cluster, in dem jene Metaboliten zu finden sind, die mit geringer Konzentration zu Beginn des Fermentationsexperimentes starten, aber auch gegen Ende wieder zu geringen Konzentrationen zurückkehren. Dieser Cluster ist sehr heterogen und enthält neben Zeitreihen, die

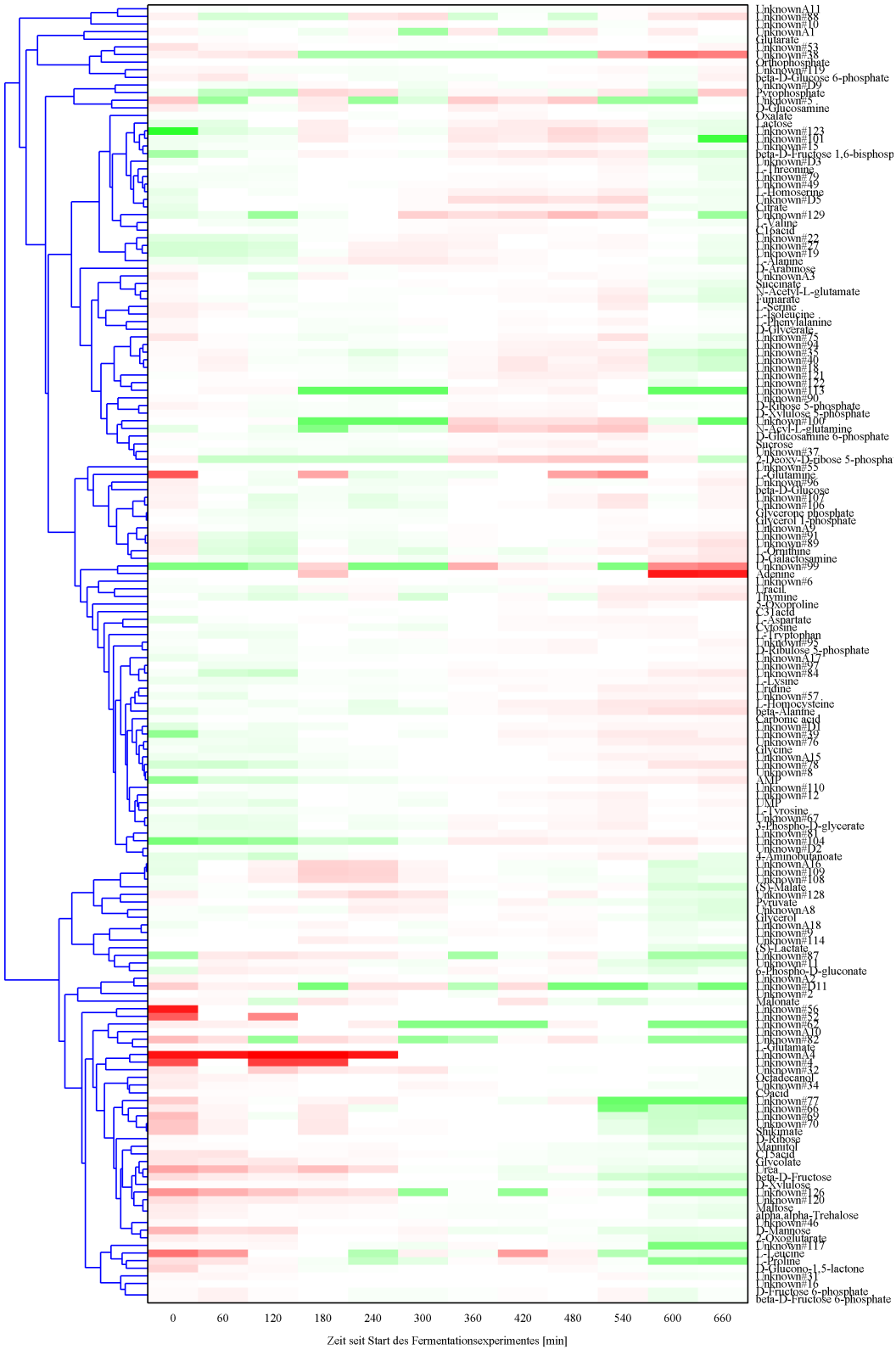

Abbildung 7.3: Clustergramm der Metaboliten der Fructose-Fermentation (050602)

ihre höchste Konzentration in der exponentiellen Phase besitzen, ferner einige zweigipflige Konzentrationsverläufe. Die exponentielle Wachstumsphase erstreckt sich nach Betrachtung der optischen Dichte (OD) von ca. 420 bis 660 Minuten. Während vor allem das Ende der exponentiellen Phase deutlich im letzten Cluster abgegrenzt wird, ist der Beginn des exponentiellen Zellwachstums nicht so deutlich in der Clustergramm-Darstellung abzuleiten, wie vergleichsweise in der Glucose- und Fructose-Fermentation.

7.1.4.2.4 Untersuchung der Lactat-Fermentation

Bei der Betrachtung der Lactat-Fermentation in Abbildung 7.5 sieht die Situation wiederum ein wenig andersartig aus. Eine grobe Einteilung in 2 Cluster ist erkennbar, wovon der erstere die Metaboliten hoher Endkonzentration enthält. Dieser Cluster erstreckt sich ungefähr vom oberen Ende der Darstellung bis zum Metaboliten 2-Oxoglutarate (C00026). Dieser erste Cluster kann jedoch weiter unterteilt werden, und zwar in solche Metaboliten, die mit niedrigen relativen Konzentrationen in die Fermentation starten - von oben bis 5-Oxoproline (C01879) - und solchen, die sowohl zu Beginn als auch zu Ende der Fermentation vergleichsweise hohe Konzentrationen aufweisen („Unknown 5“ bis 2-Oxoglutarate). Im ersten Subcluster können erneut zahlreiche Aminosäuren wiedergefunden werden. Der zweite Subcluster zeigt hingegen ein relativ heterogenes Verhalten. Von „Unknown A17“ bis L-Alanine (C00041) erstrecken sich jene Metaboliten, die in ihrem Konzentrationsverhalten eher der Form der OD-Kurve entsprechen. Dies ist gekennzeichnet durch einen Anstieg der Konzentration bis hin zum Ende der exponentiellen Wachstumsphase (welche sich ca. von 420 bis 600 Minuten erstreckt), sowie einem anschließenden Rückgang der Konzentration. Im letzten großen Cluster sind alle diejenigen Metaboliten enthalten, die tendenziell mit hohen Konzentrationen beginnen, die jedoch mehr oder weniger stetig bis zum Ende des Fermentationsexperimentes abnehmen. Zu dieser Gruppe gehört auch (S)-Lactate (C00186), welcher das Ausgangssubstrat dieser Fermentation darstellt.

7.1.4.2.5 Untersuchung der Glutamin-Fermentation

Wie bereits erwähnt, ist das Wachstum des Bakteriums *C. glutamicum* unter Anzucht mit Glutamin am geringsten im Vergleich zu allen anderen betrachteten Fermentationen. Dies ist insofern nicht verwunderlich, da alle benötigten Stoffwechselprodukte nur aus L-Glutamine (C00064), einer Aminosäure hergestellt

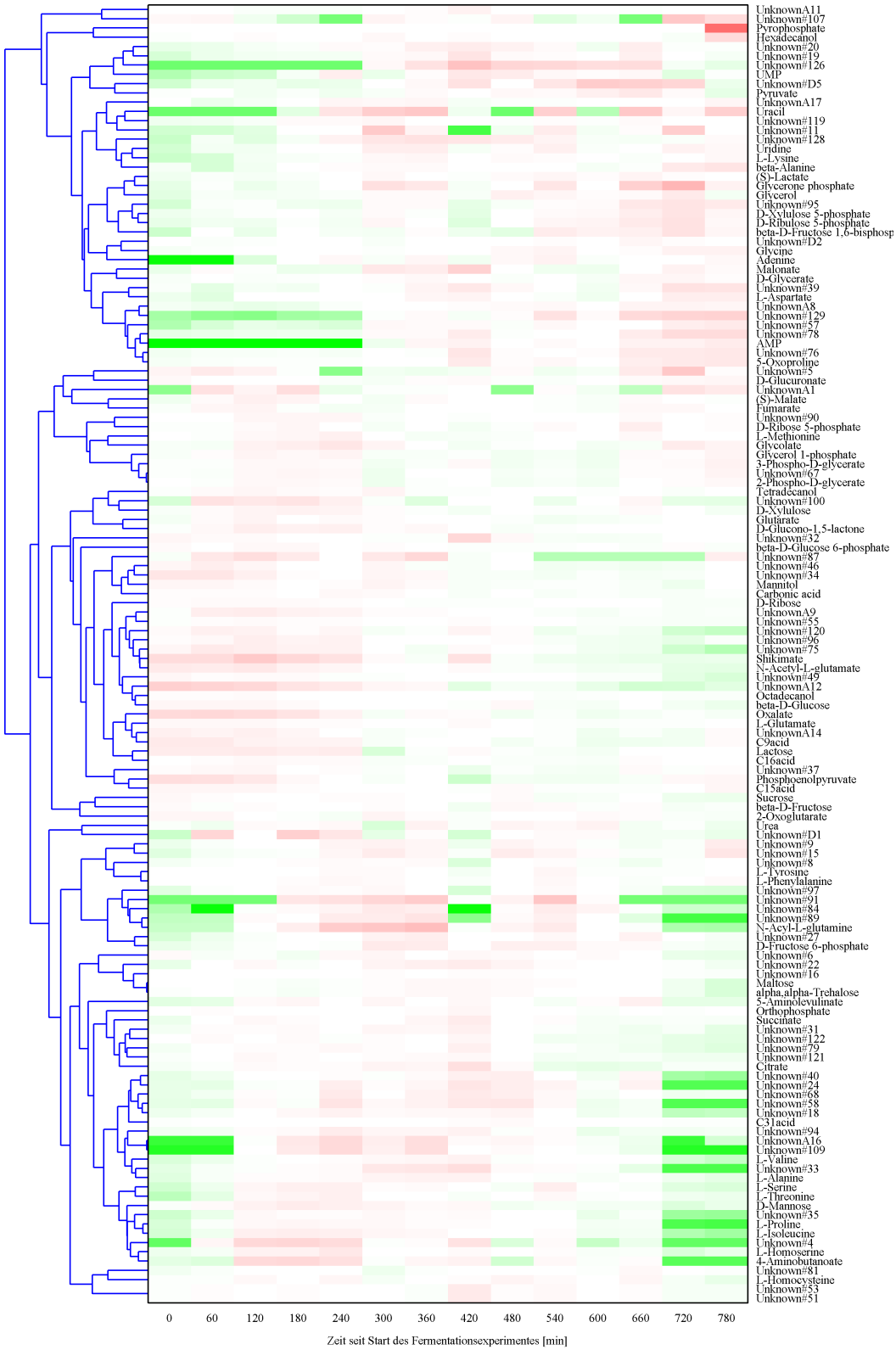

Abbildung 7.4: Clustergramm der Metaboliten der Acetat-Fermentation (050520)

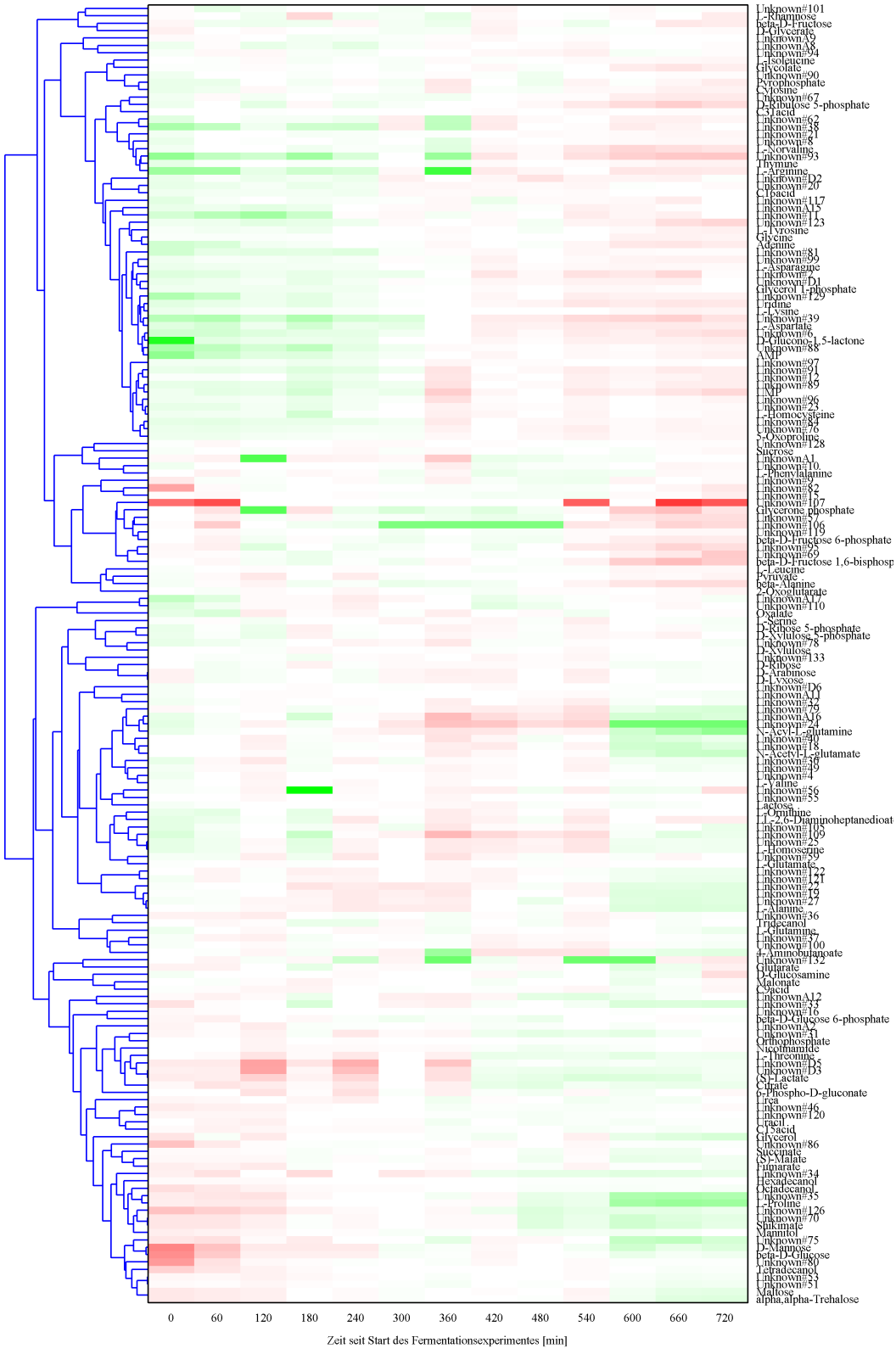

Abbildung 7.5: Clustergramm der Metaboliten der Lactat-Fermentation (050902)

werden müssen. Zum Zeitpunkt der Durchführung dieser Arbeit existierte keine wissenschaftliche Publikation, welche das Wachstum von *C. glutamicum* unter diesen Bedingungen beleuchtet. Schaut man sich daher die theoretisch gangbaren Stoffwechselwege an, so kann gemutmaßt werden, dass der Stoffffluss gänzlich über eine Umwandlung von L-Glutamine zu L-Glutamate (C00025) und von dort weitergehend zu 2-Oxoglutarate (C00026), einem Metaboliten des Zitratzyklus erfolgt. Die exponentielle Wachstumsphase erstreckt sich laut Messung der optischen Dichte von ca. 840 Minuten bis hin zum Ende des Experiments. Es zeigt sich ferner, dass sich die Anzucht auf Glutamin auch in der Clustergrammdarstellung in Abbildung 7.6 deutlich von allen anderen Fermentationen unterscheidet. Zum einen finden sich einige Cluster, in denen mehrgipfelige Konzentrationsreihen enthalten sind. Darüber hinaus ist zu erkennen, dass die Gruppe jener Metaboliten, welche mit hohen Konzentrationen beginnen und kontinuierlich in ihrer Konzentration abnehmen, vergleichsweise klein ist. Dies könnte unter Umständen eine Ursache darin haben, das ausgehend von L-Glutamine (C00064) in erster Linie Vorläufermetaboliten in ausreichender Konzentration synthetisiert werden müssen, welche im Laufe des weiteren Zellwachstums aufgebraucht werden. Eine solche Limitierung könnte ursächlich für die mehrgipfeligen Konzentrationsverläufe sein, dies jedoch sollte in weiterführenden Studien tiefergehend untersucht werden.

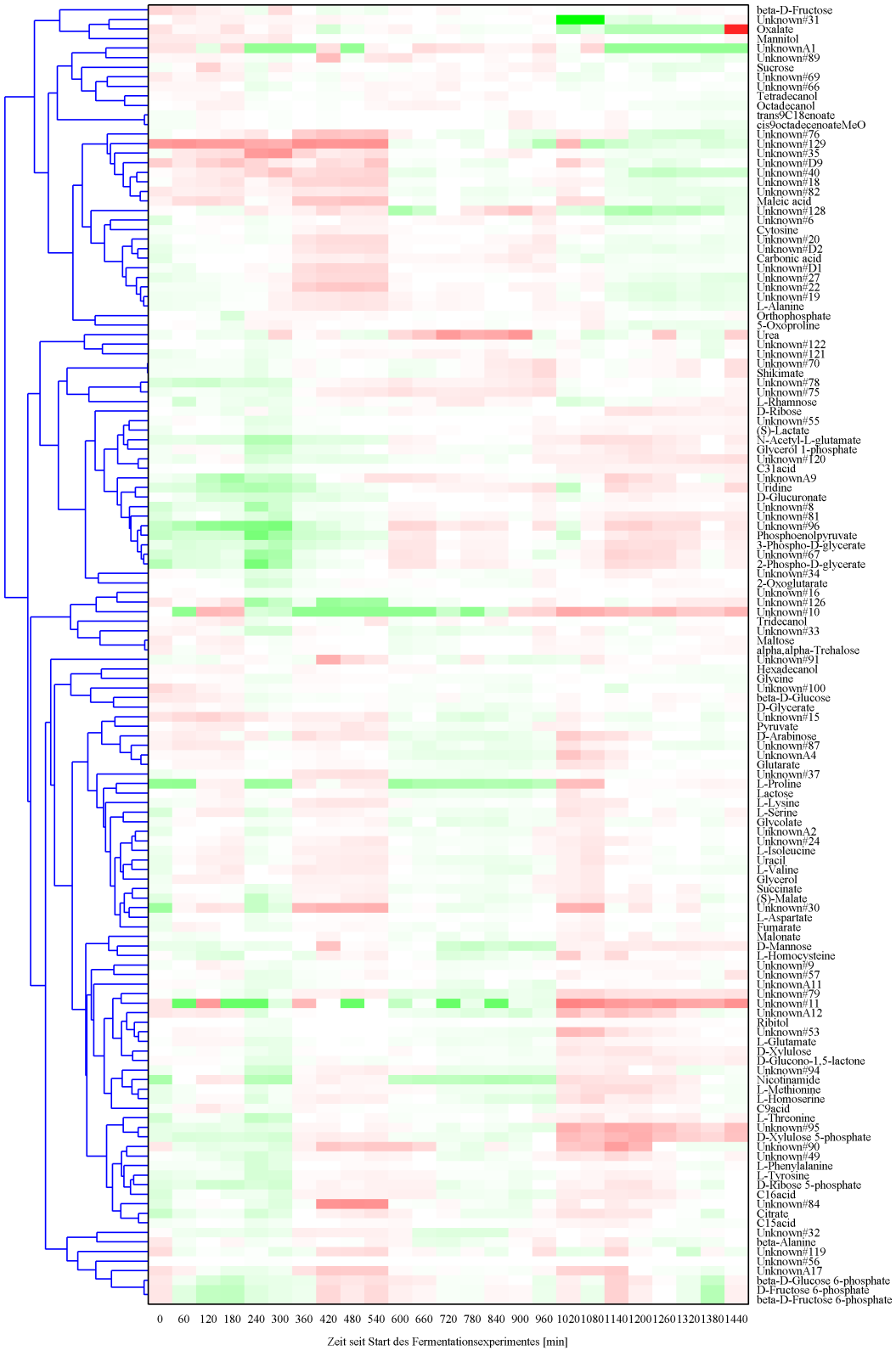

Abbildung 7.6: Clustergramm der Metaboliten der Glutamin-Fermentation (050714)

7.1.5 Gemeinsame Betrachtung aller Fermentationsexperimente

Im vorangegangenen Kapitel konnte gezeigt werden, dass deutliche Strukturmerkmale in den Zeitreihen der Metabolitkonzentrationen existieren, welche zum Teil auch fermentationsübergreifend beobachtet werden konnten. In diesem Kapitel steht hingegen eine gemeinsame Betrachtung aller Fermentationsexperimente im Vordergrund. Ein Aspekt dieses Vorgehens bestand darin, zu überprüfen, ob die Experimente hinsichtlich des gefütterten Ausgangssubstrates voneinander getrennt werden können und inwieweit sich der Stoffwechsel bei bestimmten Substraten zueinander eher ähnlich verhält.

Um diesen Vergleich durchzuführen, wurden die Konzentrationszeitreihen aller Fermentationen in einer gemeinsamen Datenmatrix zusammengeführt. Um Vergleichbarkeit zu gewährleisten, wurden die Daten gleichartig vorverarbeitet, dass heißt: ausreißer- und nullwertkorrigiert, logarithmiert und anschließend medianzentriert. Da Voruntersuchungen bereits zeigen konnten, dass bestimmte Metaboliten nicht in allen Fermentationen vorkommen, und dieser Erkenntnis hohe Wichtigkeit beizumessen ist, wurde bewusst nicht die Schnittmenge aller gleichermaßen in allen Experimenten vorkommenden Metaboliten verwendet. Ebenso wurden die nicht identifizierten Metaboliten in der Datenanalyse belassen, weil sie bis zu 30% des Datenumfanges ausmachen können.

Auf der oben beschriebenen Datengrundlage wurde eine Hauptkomponentenanalyse (PCA) berechnet. Die Grafik 7.7 zeigt, dass die substratinduzierten Unterschiede im Stoffwechsel von *C. glutamicum* deutlich zu erkennen sind. Die Fermentationsexperimente können hinsichtlich des verwendeten Ausgangssubstrates deutlich voneinander unterschieden werden. Es sind keine Überlappungen zwischen Datenpunkten verschiedener Substratgruppen festzustellen. Es kann weiterhin beobachtet werden, dass die Wiederholungsexperimente bei allen Substratgruppen sehr eng beieinander liegen. Dies spricht dafür, dass die Varianz innerhalb der Substratgruppen (trotz der zum Teil langen Zeiträume zwischen der Durchführung der Wiederholungsexperimente) deutlich geringer ist, als die substratinduzierten Unterschiede. Ferner kann auch beobachtet werden, dass sich die Glucose- und die Fructose-Fermentationen stärker ähneln als die restlichen Fermentationen. Zwischen den Substratgruppen Acetat und Lactat konnte, im Gegensatz zu ursprünglichen Annahmen keine hohe Ähnlichkeit festgestellt wer-

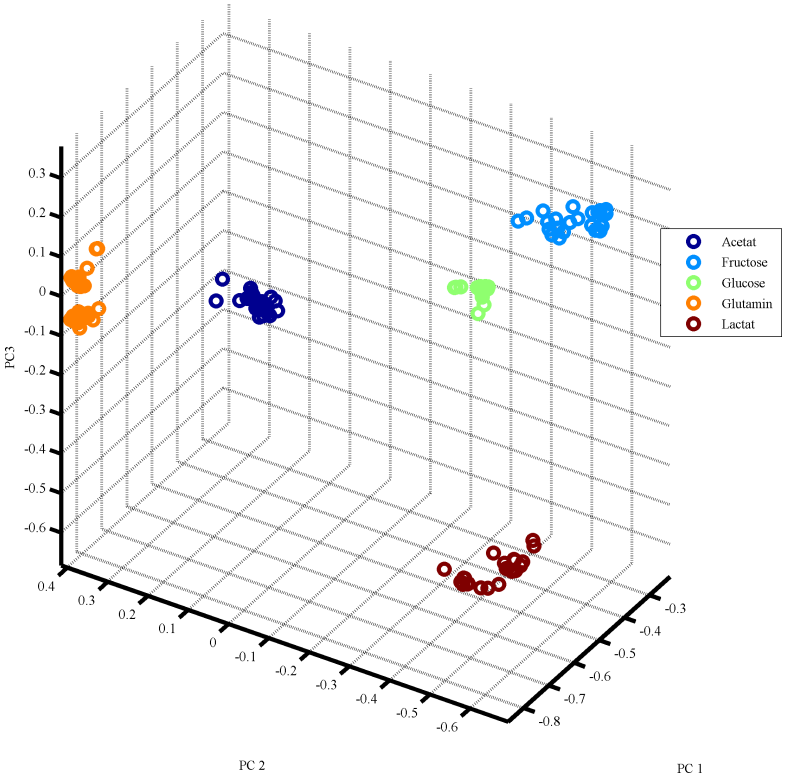

Abbildung 7.7: Hauptkomponentendarstellung der Messzeitpunkte aller Fermentationsexperimente basierend auf den ersten 3 Hauptkomponenten. Verwendete Daten: Glucose-, Fructose-, Acetat-, Lactat- und Glutamin-Fermentationen. Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung.

den. Die Glutamin-Fermentation verhält sich hingegen gänzlich anders als die restlichen Fermentationen, was auch in den Clustergramm-Darstellungen beobachtet werden konnte. Zusammengefasst zeigt die Hauptkomponentendarstellung, dass die dargebotenen Ausgangssubstrate bei ähnlichen Versuchsbedingungen zu stark unterschiedlichem temporalen Verhalten zahlreicher Metaboliten führen. Dies ist als Hinweis darauf zu werten, dass unterschiedliche Stoffwechselwege in *C. glutamicum* unter den betrachteten Bedingungen aktiv sein müssen.

7.2 Analyse der theoretischen Ausgangsdaten

Wie bereits ausführlich behandelt, dienten Reaktionsnetzwerke als Grundlage für die weitere Analyse. Es lagen insgesamt vier verschiedene Reaktionsnetzwerke vor, die aus der Bielefelder Annotation (CGB), der Kyowa Hakko-Annotation (CGL) sowie in den Varianten VGL1 und VGL2 aus der durchgeführten Genomannotation sowie unter Hinzunahme von weiterführender Information hervorgegangen sind. Das Pathway Hunter Tool (PHT) wurde verwendet, um eine Suche nach biochemisch sinnvollen Stoffwechselwegen innerhalb der betrachteten Reaktionsnetzwerke durchzuführen. Darüber hinaus wurden weiterführende beschreibende Größen aus den vom PHT gelieferten Informationen abgeleitet: die Topologie-Deskriptoren (vergleiche hierzu insbesondere Kapitel 5.2.2).

Um einen Überblick zu geben, wird in den nachfolgenden Kapiteln zuerst auf die grundlegenden Eigenschaften metabolischer Netzwerke eingegangen. Diese werden anhand einer graphischen Darstellung des metabolischen Netzwerkes von *C. glutamicum* erläutert.

Auf dieses Kapitel folgt eine detaillierte Betrachtung, welche die Unterschiede zwischen den oben genannten Reaktionsnetzwerken und den daraus resultierenden Modellierungen zum Ziel hat. Um diesen Vergleich durchführen zu können, wurden vordefinierte Metabolitkombinationen für die Modellierung der Stoffwechselwege herangezogen (vergleiche Kapitel 4.3.3). Inhaltlich beschäftigt sich dieses Kapitel beispielsweise mit der Frage, welche zusätzlichen Stoffwechselwege unter Verwendung des Reaktionsnetzwerkes VGL1 gefunden werden konnten. Eine andere Frage, die in diesem Zusammenhang untersucht wurde ist, inwiefern sich die Wahl des Mapping-Algorithmus auf die Anzahl gefundener Pfade auswirkt.

Um eine einheitliche Nomenklatur sicherzustellen, wird im weiteren Verlauf

dieser Arbeit die Modellierung, welche auf dem Reaktionsnetzwerk der Bielefelder Annotation von *C. glutamicum* basiert, „CGB-Modellierung“ genannt, während für die Modellierung basierend auf der Kyowa Hakko- Annotation der Begriff „CGL-Modellierung“ verwendet wird. Entsprechend werden die Modellierungen auf den neuen Reaktionsnetzwerken analog dazu „VGL1-Modellierung“ und „VGL2-Modellierung“ genannt.

7.2.1 Grundlegende Betrachtung metabolischer Netzwerke

Metabolische Netzwerke können mit Hilfe graphentheoretischer Ansätze gut visualisiert werden. Häufig wird die Darstellung dergestalt gewählt, dass die Metaboliten die Knotenpunkte und die enzymatischen Reaktionen die Verbindungslinien des Netzwerkes repräsentieren. In diesem Falle spricht man von einer metabolitzentrischen Betrachtungsweise. Der umgekehrte Fall ist auch möglich, in diesem Fall spricht man von einer enzym-zentrischen Darstellung. Betrachtet man ein metabolit-zentrisch dargestelltes metabolisches Netzwerk, wie zum Beispiel das metabolische Netzwerk von *C. glutamicum* in Abbildung 7.8, so können folgende grundlegenden Eigenschaften festgestellt werden:

- Metabolische Netzwerke sind im Allgemeinen nicht vollständig konnektiert, das heißt, es existieren isolierte Subnetze (Ma und Zeng, 2003b). Dies bedeutet, dass (vom theoretischen Standpunkt her) nicht jeder beliebige Metabolit in einen anderen umgesetzt werden kann. Die Ursache hierzu kann unter Umständen mit der Tatsache verknüpft sein, dass noch nicht alle katalysierenden Enzyme entdeckt worden sind.

- Neben einem großen, vollständig konnektierten Teil des Netzwerkes exisitieren viele kleine isolierte Subnetze, welche nur aus wenigen Metaboliten bestehen. Vollständig konnektierte Strukturen innerhalb von Netzwerken werden in der Graphentheorie „strong component“ genannt (Batagelj und Mrvar, 1998) und konnten neben metabolischen Netzwerken (Ma und Zeng, 2003a) auch in zahlreichen anderen Formen von Netzwerken (beispielsweise des Internets) festgestellt werden. In Abbildung 7.8 ist - rot eingefärbt - die größte zusammenhängende Netzwerkstruktur gekennzeichnet. Da es sich bei der Darstellung um einen gerichteten Graphen handelt, sind alle Metaboliten zwar miteinander verbunden, aber nicht reversibel in einander

umsetzbar. Möchte man nur jene Metaboliten betrachten, die reversibel ineinander umgesetzt werden könnten, so würde des lediglich einem Teil der vollständig konnektierten Struktur entsprechen.

- Es ist deutlich zu sehen, dass die Mehrzahl der Metaboliten nur wenige Nachbarn besitzt, während einige, wenige Metaboliten stark verknüpft sind. Diese Auffälligkeit im Verknüpfungsverhalten findet sich in zahlreichen Arten von Netzwerken und kann durch die Potenz-Verteilungsfunktion (engl. Power Law Distribution) beschrieben werden (Jeong et al., 2000; Palsson et al., 2003). Die wenigen, stark im Netzwerk verknüpften Metaboliten werden „Hubs“ genannt und sind maßgeblich für die Struktur des gesamten Netzwerkes (Bray, 2003). In metabolischen Netzwerken haben sie oft eine essentielle biochemische Funktion inne und sind häufig im Zentralstoffwechsel anzusiedeln. Typische Beispiele für metabolische Netzwerk-Hubs aus dieser Arbeit sind beispielsweise die Metaboliten Pyruvate, L-Glutamate oder 2-Oxoglutarate (vergleiche hierzu insbesondere Kapitel 7.4.1.1). Untersuchungen von Stelling et al. (2002), Schilling et al. (2002) oder Ravasz et al. (2002) konnten zeigen, dass in den metabolischen Netzwerken zahlreicher Organismen identische Metaboliten als Netzwerk-Hubs identifiziert werden konnten.

- Ohne die ausdrückliche Eliminierung von Reaktionswegen über Seitenmetaboliten (vergleiche hierzu insbesondere Kapitel 4.3.3) ist die graphische Darstellung metabolischer Netzwerke nahezu unbrauchbar.

7.2.2 Detaillierte Betrachtung metabolischer Netzwerke

Für die detaillierte Untersuchung der Reaktionsnetzwerke wurde - wie bereits erwähnt - das in Kapitel 4.3.3 beschriebene Vorgehen gewählt, wobei das Pathway Hunter Tool (PHT) annotationsspezifisch mit der Suche von insgesamt 15006 potenziellen Pfadkombinationen gestartet wurde. Diese vordefinierten Pfadkombinationen beruhen auf Paarungen zwischen Metaboliten, die bereits experimentell in *C. glutamicum* erfasst werden konnten und stellen folglich eine Teilmenge des gesamten Netzwerkes dar. Dieser Schritt wurde bewusst gewählt, da in der integrativen Analyse ohnehin nur die Schnittmenge gemeinsam erfasster Informationen analysiert werden kann. Ferner konnte somit eine deutliche Verminderung der

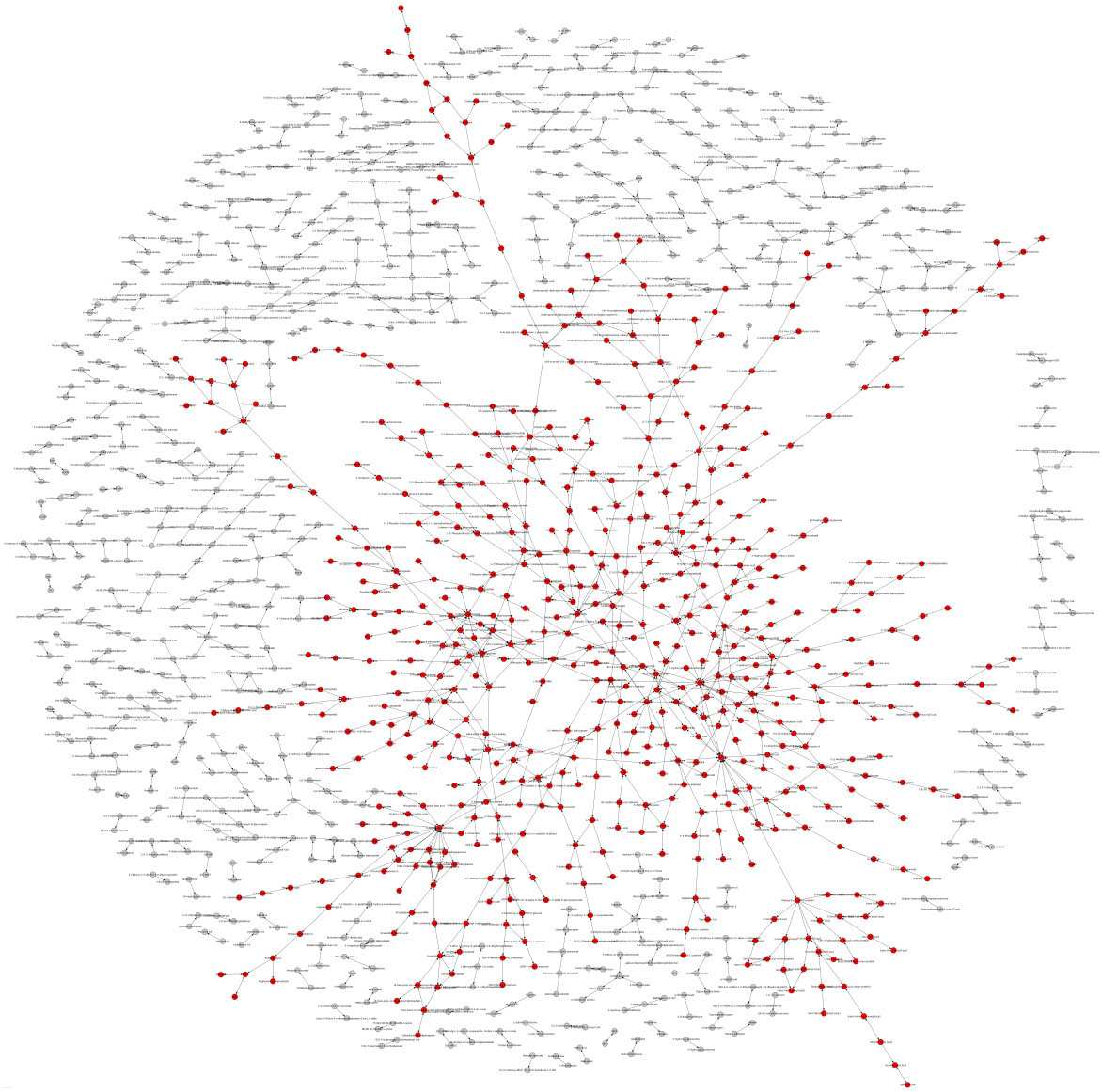

Abbildung 7.8: Schematische graphische Darstellung des theoretischen metabolischen Netzwerkes von *C. glutamicum*. Als Reaktionsnetzwerk für die Modellierung wurde das aus der Annotation abgeleitete VGL1-Netzwerk verwendet. Der KEGG-Mapping-Algorithmus wurde verwendet sowie die Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen. Die Richtungsabhängigkeit von Reaktionen ist mit Indikatoren dargestellt. Rot eingefärbte Metaboliten stellen die größte konnektierte Struktur dar. Isolierte Subnetzwerke sind grau eingefärbt. Die Erstellung des Graphen erfolgte mit dem yEd Graph Editor (`www.yworks.com`).

für die Modellierungen benötigten Rechenzeit erreicht werden. Weiterhin ermöglichte es, eine höhere Anzahl beschreibender Deskriptoren auf dem reduzierten Datensatz zu berechnen.

Die wichtigste Kennzahl im Vergleich der Netzwerkmodellierungen ist die Anzahl gefundener gültiger metabolischer Pfade. Die nachfolgende Tabelle 7.8 gibt einen ersten Überblick darüber, wie viele gültige Pfade bei den gegebenen Voraussetzungen gefunden werden konnten. Erste generelle Aussagen lassen sich wie folgt zusammenfassen:

- Selbst für das betrachtete Subset experimentell erfasster Metabolitpaarungen konnten bei weitem nicht für alle Kombinationen auch tatsächlich gültige Pfade gefunden werden. Die Ursache hierfür liegt vermutlich in der bereits in Kapitel 7.2.1 festgestellten nicht vollständigen Konnektivität metabolischer Netzwerke. Diese Erkenntnis steht in Übereinstimmung zu anderen Untersuchungen metabolischer Netzwerke (Csete und Doyle, 2004).

- Die Anzahl gefundener Pfade unterscheidet sich deutlich in Abhängigkeit der verwendeten Reaktionsnetzwerke als auch des verwendeten Mapping-Algorithmus.

- Die Verwendung der neuen Reaktionsnetzwerke VGL1 und VGL2 führte zu einer deutlich gesteigerten Anzahl gefundener Pfade. Die Zunahme konnte sowohl unter KEGG- als auch unter CUBIC-Bedingungen beobachtet werden. So beträgt beispielsweise der Zuwachs von CGB zu VGL1 unter KEGG-Bedingungen ca. 46% während sie unter CUBIC-Bedingungen ca. 20% beträgt. Die Erhöhung der Anzahl gefundener Pfade ist in erster Linie dem erweiterten Wissen über den Enzymkatalog des Organismus (vergleiche Kapitel 4.3.1) zurückzuführen und kann folglich nachvollzogen werden.

- Vergleicht man die Mapping-Algorithmen, so führt die Verwendung des CUBIC-Mappings im Vergleich zu KEGG-Bedingungen generell zu einer höheren Anzahl gefundener Pfade. Der durch die Verwendung des Mapping-Algorithmus induzierte Effekt fällt bei den CGB- und CGL-Modellierungen mit einer Steigerung von rund 52% im Vergleich zum KEGG-Mapping sehr deutlich aus, während bei den VGL1- und VGL2-Modellierungen eine Steigerung von 25% erreicht werden kann. Als Erklärung ist eventuell zu sehen,

dass weniger stark verknüpfte Netzwerke deutlich sensitiver auf das CUBIC-Mapping reagieren. Festzustellen ist allerdings jedoch, dass der Effekt des CUBIC-Mappings nicht immer eindeutig nachvollzogen werden kann.

Eine detaillierte Betrachtung der einzelnen Modellierungen, ihrer Unterschiede sowie den vermutlich zugrunde liegenden Ursachen erfolgt in den nächsten Unterkapiteln.

Tabelle 7.8: Tabellarischer Vergleich der vier in dieser Arbeit untersuchten Reaktionsnetzwerke für *C. glutamicum*. Darstellung der Anzahl zugrunde liegender Enzyme, der Anzahl vorhandener Metaboliten und Reaktionen, sowie der Anzahl von metabolischen Pfaden, die mit dem PHT unter KEGG- und CUBIC-Bedingungen gefunden werden konnten.

		CGB	CGL	VGL1	VGL2
	Anzahl Enzyme	554	538	604	668
	Anzahl Metaboliten	1069	1075	1557	1604
	Anzahl Reaktionen	907	899	1435	1520
KEGG:	Pfade gefunden	2559	2543	3725	3901
CUBIC:	Pfade gefunden	3903	3862	4682	4902

Generell muss in diesem Zusammenhang erwähnt werden, dass es sich bei der Modellierung der metabolischen Pfade um eine rein theoriebasierte Betrachtungsweise des Stoffwechsels von *Corynebacterium glutamicum* handelt. Die rechnergestützte Suche nach Stoffwechselwegen beruht hierbei auf umfangreichem Wissen wie beispielsweise der in einem Organismus vorhandenen Enzyme und der daraus resultierenden biochemischen Reaktionen. Die Suche nach gültigen Stoffwechselwegen zwischen zwei Metaboliten ist daher ein theoretisches Konstrukt, welches gewissermaßen aus sehr viel umfangreicherem Wissen abgeleitet wurde. Ferner sagt ein durch informatische Werkzeuge gefundener Stoffwechselweg lediglich aus, dass dieser Weg bei dem momentanen Stand der Wissenschaft organismenspezifisch höchstwahrscheinlich existiert und auch gangbar ist. Wie stark der Stoffwechselweg allerdings in der Realität frequentiert ist und wie wichtig er für den Organismus als solchen ist, lässt sich verständlicherweise aus dieser Betrachtung nicht ableiten.

7.2.2.1 Vergleich der Mapping-Verfahren

Unter dem Begriff „Mapping-Verfahren“ sind Regeln zusammengefasst, die reaktionsspezifisch definieren, welche Metaboliten in welcher Richtung ineinander umgesetzt werden können. Oder anders ausgedrückt, sie beschreiben, welche Verbindungen zwischen Metaboliten reaktionsspezifisch erlaubt sind. In Kapitel 5.2.4.1 wurde die Thematik bereits in einem anderen Zusammenhang angerissen. Besteht eine chemische Reaktion auf beiden Seiten der Gleichung aus mehreren Reaktionspartnern (was meistens der Fall ist), so muss geklärt werden, welche Edukte in welche Produkte überführt werden können. Mit dem KEGG- und dem CUBIC-Mapping wurden zwei verschiedene Ansätze getestet. Das KEGG-Mapping orientiert sich eng am Prinzip der KEGG-Pathway Maps und beinhaltet daher Informationen, wie sie von zahlreichen Forschergruppen zusammengetragen worden sind.

Der von Dr. Syed Asad Rahman entwickelte CUBIC-Algorithmus versucht zusätzliche, reaktionsspezifisch gültige Kombinationen zwischen Metaboliten unter Berücksichtigung ihrer molekularen Struktur (auch unter anderen beteiligten Reaktionspartnern) zu finden (persönliche Kommunikation mit Dr. S. A. Rahman). In anderen Worten ausgedrückt: das CUBIC-Mapping ist dadurch charakterisiert, dass es im Vergleich zum KEGG-Mapping eine deutlich höhere Anzahl von reaktionsspezifisch gültigen Verknüpfungen zwischen Metaboliten erlaubt. Dies resultiert im Allgemeinen in einer insgesamt höheren Anzahl gefundener Pfaden sowie einer im Mittel kürzeren Pfadlänge. Der Unterschied zwischen den beiden Mapping-Verfahren wird nachfolgend exemplarisch anhand der Bielefeld-Modellierung exemplarisch erörtert.

Wie die Tabelle 7.8 gezeigt hat, wurden bei der CGB-Modellierung unter Verwendung des CUBIC-Mappings 3903 metabolische Pfade gefunden, während unter KEGG-Bedingungen mit 2559 nur rund 65% des Umfangs erreicht wurden. Die Schnittmenge der bei beiden Mapping-Algorithmen gleichermaßen gefundenen Pfadkombinationen liegt bei 2462. Bei genauerer Betrachtung bedeutet dies interessanterweise, dass 97 individuelle Pfadkombinationen nur unter KEGG-Bedingungen gefunden werden konnten. Dies bezieht sich ausschließlich auf Pfade, welche auf die Metaboliten LL-2,6-Diaminoheptanedioate (C00666), meso-2,6-Diaminoheptanedioate (C00680), L-Lysine (C00047) und Uracil (C00106) enden. Eine detaillierte Betrachtung der Ausgabedatei des PHT ergab, dass für diese

Pfadkombinationen chemische Strukturinformationen eines oder mehrerer Metaboliten innerhalb des Pfades fehlten und deshalb entweder die Ähnlichkeitsberechnung nicht durchgeführt werden konnte, oder die Pfade die gewählten Ähnlichkeitskriterien nicht erreichten. Für 1441 Kombinationen gilt das Gegenteil, sie konnten nur unter CUBIC-Bedingungen ermittelt werden. Wenn man die gemeinsame Schnittmenge der 2462 Pfadkombinationen betrachtet, sind die Pfadlängen der gefundenen metabolischen Pfade unter KEGG- und CUBIC-Bedingungen oft nicht identisch. Für rund die Hälfte der gemeinsamen Kombinationen (1283 von 2462) konnten Unterschiede hinsichtlich der Pfadlänge festgestellt werden.

Die Abbildung 7.9 zeigt exemplarisch am Beispiel der Bielefeld-Modellierung, wie die Pfadlänge durch die Wahl des Mapping-Algorithmus beeinflusst wird. In allen betrachteten Abweichungen sind die mit dem CUBIC-Mapping ermittelten Pfade kürzer als die Entsprechungen aus dem KEGG-Mapping. Vereinzelt konnten auch Extrema beobachtet werden, so beträgt die größte Differenz in der Pfadlänge 16 Reaktionsschritte, was sich auf den metabolischen Pfad von N-Acetyl-L-Glutamate (C00624) zu LL-2,6-Diaminoheptanedioate (C00666) bezieht. Unter Verwendung des KEGG-Mappings ist die Reaktionskette 19 Schritte lang, während unter CUBIC-Bedingungen ein nur 3 Schritte langer Umsetzungweg über L-Glutamate (C00025) und den Lysin-Stoffwechsel berechnet wird. Die geringste Differenz in der Pfadlänge beträgt nur einen Reaktionsschritt, die mittlere Abweichung liegt bei 3,9 Reaktionsschritten. Eine ausführliche Betrachtung ergab, dass die Mehrheit der Unterschiede mit geringen Differenzen in der Pfadlänge einhergeht. Die Konsequenz aus den induzierten Unterschieden ist, dass sich neben der Pfadlänge auch die weiteren aus dem PHT-Output angeleiteten Deskriptoren (vergleiche Kapitel 5.2.2) unterscheiden. Erneut wird wieder deutlich, dass die Modellierung von Pfaden innerhalb eines metabolischen Netzwerkes auch auf Annahmen beruht und somit eine rein theoretische Betrachtungsweise darstellt. Wie stark die ermittelten Pfade in der Realität frequentiert sind, kann diese theoretische Betrachtung nicht beantworten.

7.2.2.2 Vergleich der Bielefelder- mit der Kyowa Hakko-Modellierung

7.2.2.2.1 Unter KEGG-Bedingungen

Da die beiden Annotationen Unterschiede in ihrem Umfang als auch in der Anzahl und Zusammensetzung der annotierten Gene aufweisen, spiegelt sich dieser

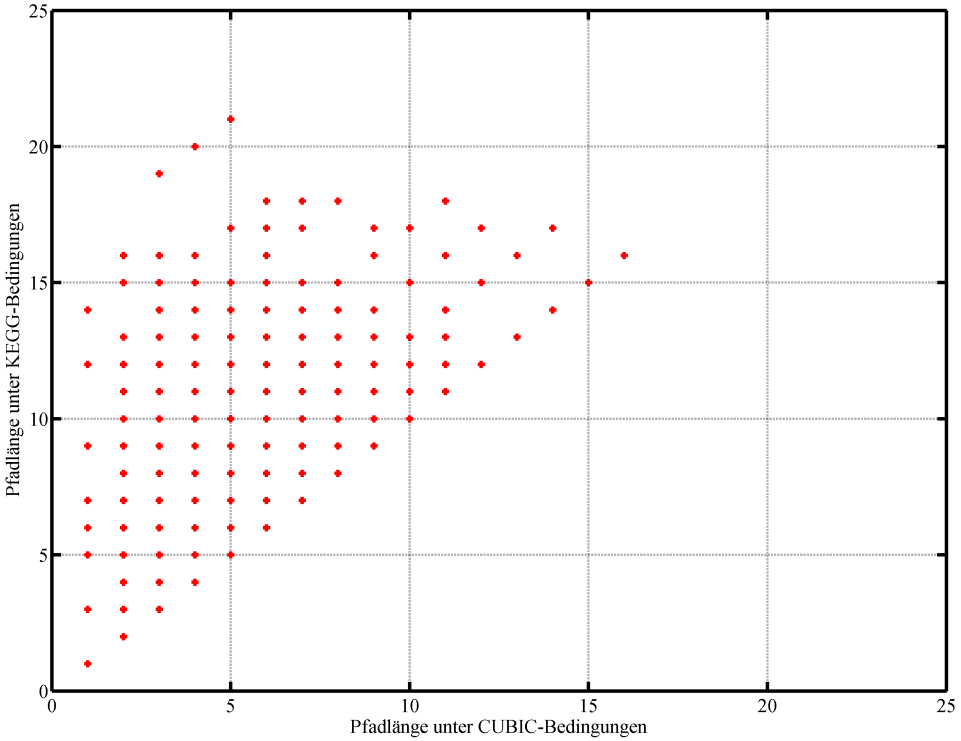

Abbildung 7.9: Einfluss des Mapping-Algorithmus auf die Pfadlänge. Als Reaktionsnetzwerk für die Modellierung wurde das aus der Bielefelder Annotation für *C. glutamicum* abgeleitete Reaktionsnetzwerk verwendet. Das KEGG- und CUBIC-Mapping sowie die Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit wurden verwendet. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen.

Sachverhalt auch in der Modellierung der metabolischen Pfade wider. So finden sich beispielsweise bei der Bielefeld-Modellierung keine Pfade, die vom Metaboliten D-Glucono-1,5-lactone (C00198) ausgehen. Eine Betrachtung der PHT-Programmausgabe ergab, dass dieser Metabolit für jene Modellierung nicht vorhanden ist. Ein detaillierter Blick in die Programmausgabe des PHT bestätigt, dass das Enzym Glucose 1-dehydrogenase (EC 1.1.1.47), welches den Metaboliten D-Glucono-1,5-lactone, mit beta-D-Glucose (C00221) verbindet, sehr wohl in der Kyowa Hakko-Annotation vorkommt, aber in der Bielefelder Annotation nicht existiert. Im umgekehrten Fall existiert der Metabolit Cytosine (C00380) nicht in der Kyowa Hakko-Modellierung, wohingegen er in der Bielefelder Annotation, katalysiert durch das Enzym Cytosine Deaminase (EC 3.5.4.1), eine Reaktion mit dem Metaboliten Uracil (C00106) eingehen kann.

Neben dem Vorhandensein beziehungsweise dem Nichtvorhandensein bestimmter Enzyme in den Annotationen, können auch die bei der Modellierung gesetzten Schwellenwerte der globalen und lokalen Ähnlichkeit beim Finden der metabolischen Pfade für Unterschiede sorgen. So existieren zum Beispiel in der Kyowa Hakko-Modellierung zwischen Glycolate (C00160) und (S)-Lactate (C00186) keine metabolischen Pfade, welche allerdings unter identischen Bedingungen (lokale Ähnlichkeit 15%, globale Ähnlichkeit 1% vergleiche Kapitel 4.3.3) in der Bielefeld-Modellierung gefunden werden konnten. Gleiches betrifft ferner einige Pfade, die von Maltose (C00208) ausgehen. Die Ursache hierfür kann unter Umständen im Nichtvorhandensein einiger molekülspezifischen Fingerprints entlang des Pfades zu suchen sein, wodurch die Ähnlichkeitsberechnung und damit die Pfadsuche verhindert wird.

Betrachtet man die Schnittmenge der in beiden Modellierungen gemeinsamen Kombinationen, so sind dies 2443 gemeinsame Pfade, was ausgehend von der Bielefeld-Modellier- ung einem sehr hohen Anteil von ca. 95% Prozent entspricht. Vergleicht man analog auch hier die Pfadlängen der gemeinsam vorhandenen Kombinationen, so ergibt sich folgendes Bild. Für annähernd alle in beiden Modellierungen gefundenen Pfade war die Pfadlänge identisch. Es konnten lediglich in 46 von 2443 Fällen Unterschiede in der Pfadlänge festgestellt werden, was nur knapp 2% der gemeinsamen Kombinationen betrifft. Diese Unterschiede traten lediglich bei metabolischen Pfaden auf, welche von Sucrose (C00089) zu anderen Metaboliten ausgehen. Die betroffenen Pfade waren in der Bielefeld-Modellierung um einen einzigen Reaktionsschritt kürzer. Um die Ursache für diese systematischen Unter-

schiede zu ermitteln, wurde der kürzeste metabolische Pfad, bei dem dieser Effekt noch auftrat, untersucht. So ist beispielsweise der Pfad zwischen Sucrose und beta-D-Fructose 1,6-bisphosphate (C05378) in der Bielefeld-Modellierung 3 Schritte lang, während in der Kyowa Hakko-Modellierung 4 Schritte notwendig sind. Der Grund dafür liegt darin, dass das Enzym Alpha-glucosidase (EC 3.2.1.20), welches eine direkte Umwandlung von Sucrose zu D-Fructose (C10906) ermöglicht (siehe Reaktion R00801), nicht in der Kyowa Hakko-Annotation vorhanden ist. Die mittlere Länge aller gefundenen Pfade beträgt für die CGB-Modellierung unter KEGG-Bedingungen 7,76 und für die CGL-Modellierung 7,77 Reaktionsschritte (basierend auf dem beschriebenen Setup, vergleiche Kapitel 4.3.3).

7.2.2.2.2 Unter CUBIC-Bedingungen

Da der Metabolit D-Glucono-1,5-lactone, wie bereits in Kapitel 7.2.2.2.1 angemerkt, nicht in der Bielefelder Modellierung vorhanden ist und Cytosine in der Kyowa Hakko-Modellierung fehlt, fehlen logischerweise auch unter CUBIC-Bedingungen sämtliche Pfadkombinationen, die diese Metaboliten benutzen. Zusätzliche Unterschiede im Vorhandensein der Metaboliten sind nicht zu verzeichnen.

Da deutlich mehr kombinatorische Verknüpfungsmöglichkeiten zwischen Metaboliten unter CUBIC-Bedingungen erlaubt sind, ist auch die Schnittmenge der in beiden Modellierungen gleichermaßen gefundenen Pfade größer. Die Schnittmenge zwischen der Bielefeld-Modellierung und der Kyowa Hakko-Modellierung beträgt unter CUBIC-Bedingungen 3738 Pfade, was ebenfalls einem sehr hohen Anteil von ca. 96% ausgehend von der Bielefeld-Modellierung entspricht. Allerdings fällt hier der Vergleich zwischen den jeweiligen Pfadlängen deutlich heterogener aus. Unter KEGG-Bedingungen betrug der Unterschied in der Pfadlänge nur einen Reaktionsschritt, welcher ursächlich im Vorhandensein eines Enzyms begründet ist. Bei den CUBIC-Modellierungen findet sich ein breites Spektrum an Unterschieden. Die Unterschiede in der Pfadlänge treten hierbei häufiger auf. Sie konnten in 510 der 3738 gemeinsamen Kombinationen festgestellt werden, was einem Anteil von 13,6% entspricht.

In 361 dieser 510 Kombinationen (entspricht ca. 70%) findet sich der kürzere Pfad in der CGB-Modellierung. Der größte gefundene Unterschied in der Pfadlänge beträgt unter diesen Bedingungen 6 Reaktionsschritte. So ist der Pfad zwischen den Metaboliten Uracil (C00106) und alpha-D-Glucose (C00267) in

der CGL-Modellierung 13 Reaktionsschritte lang, während dieser in der CGB-Modellierung nur 7 Schritte beträgt. Der größte Pfadunterschied, bei dem die CGL-Modellierung die kürzere Entsprechung aufweist, findet sich im Pfad von L-Alanine (C00041) zu Citrate (C00158). Dieser Weg ist in der CGB-Modellierung 7 Reaktionsschritte lang, während er in der CGL-Modellierung nur 3 Schritte lang ist. Die verursachende Abkürzung wird durch das Enzym Citrate Lyase (EC 4.1.3.6) verursacht. Obwohl dieses Enzym nachweislich in beiden Annotationen vorhanden ist, fehlt der abkürzende Reaktionsschritt (R00362) im Reaktionsnetzwerk der Bielefelder Annotation.

Die mittlere Abweichung in der Pfadlänge beträgt unter CUBIC-Bedingungen zwischen der CGB- und der CGL-Modellierung 1,58 Reaktionsschritte. Die mittlere Länge der in beiden Modellierungen gleichermaßen gefundenen Pfadkombinationen, beträgt für die CGB-Modellierung 6,05 und für die CGL-Modellierung 6,18 Reaktionsschritte. Abbildung 7.10 verdeutlicht die Unterschiede in der Pfadlänge zwischen den beiden Modellierungen und Mapping-Verfahren.

7.2.2.3 Betrachtung der VGL1-Modellierung

Um die Übersichtlichkeit zu wahren und der Tatsache Rechnung zu tragen, dass sich die Bielefelder- und die Kyowa Hakko- Annotation doch recht stark ähneln, wurden an dieser Stelle nur die Unterschiede zwischen der gemeinsamen Schnittmenge dieser beiden Modellierungen mit der VGL1-Modellierung eingehend betrachtet. Analog zum obigen Vorgehen wurden das Vorhandensein von Pfaden, sowie deren Länge unter KEGG- und CUBIC-Bedingungen untersucht.

7.2.2.3.1 Unter KEGG-Bedingungen

Die gemeinsame Schnittmenge der Bielefelder- und der Kyowa Hakko-Modellierung beträgt wie in Kapitel 7.2.2.2.1 beschrieben 2443 metabolische Pfade. Die Netzwerkmodellierung basierend auf VGL1 liefert unter KEGG-Bedingungen 3725 gültige Pfadkombinationen für die vordefinierten PHT-Startparameter. Schaut man sich die Unterschiede im Detail an, ist festzustellen, dass sämtliche Pfade aus der gemeinsamen Schnittmenge auch in der neuen Modellierung gefunden werden können. Zusätzlich liefert die VGL1-Modellierung jedoch 1282 neue Pfadkombinationen.

Für einige Metaboliten, die weder in der CGB- noch in der CGL-Modellierung

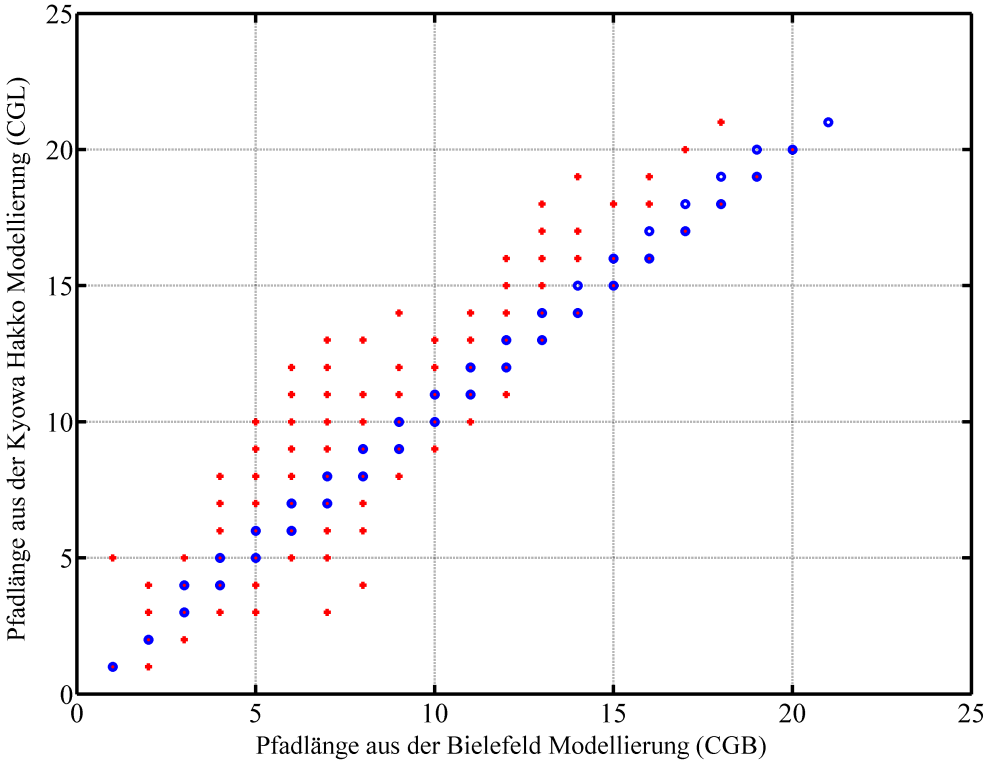

Abbildung 7.10: Vergleich der Pfadlängen der gleichermaßen in der CGB- und CGL-Modellierung gefundenen metabolischen Pfade. Betrachtung unter KEGG-Bedingungen (blau) und CUBIC-Bedingungen (rot). Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit wurden verwendet. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen.

konnektiert waren, oder nur in einer der beiden Modellierungen überhaupt erreicht wurden, konnten nun Pfadkombinationen gefunden werden. Hierzu gehören beispielsweise Pfade, die von D-Glucono-1,5-lactone (C00198), Glycolate (C00160), oder alpha,alpha-Trehalose (C01083) ausgehen oder auf Metaboliten wie D-Fructose 1,6-bisphosphate (C00354), oder Maltose (C00208) enden. Neben gänzlich neu hinzugekommenen Pfaden, wurden zahlreiche, bereits existierende Pfade in ihrer rückwärtigen Richtung erschlossen. Hierzu gehören beispielsweise auch Pfade, welche von Endprodukten des Stoffwechsels ausgehen. Als Beispiel kann der Pfad von L-Homoserine (C00263) ausgehend zu Citrate (C00158) angeführt werden, welcher zuvor nicht gefunden werden konnte. Das Finden zahlreicher Rückreaktionen deutet darauf hin, dass durch die Erweiterung des enzymatischen Repertoires zusätzliche Stoffwechselwege erschlossen werden konnten.

Beim Vergleich der Pfadlängen der, gleichsam in der Schnittmenge von Bielefelder-, Kyowa Hakko-und VGL1-Modellierung gefundenen Kombinationen, konnten Auffälligkeiten festgestellt werden. Durch die Integration zusätzlicher Enzyme konnten manche Pfade in ihrer Länge nicht unerheblich abgekürzt werden. In 416 der 2443 gemeinsam betrachteten Pfade lieferte die neue Modellierung kürzere Pfade, was einem Anteil von 17% aller betrachteten Fälle entspricht (vergleiche hierzu die Abbildung 7.11). Die mittlere Differenz in der Pfadlänge beträgt unter diesen Bedingungen 2,89 Schritte. Die maximale Differenz in der Pfadlänge beträgt 14 Schritte und betrifft den Pfad von Sucrose (C00089) zu D-Fructose 6-phosphate (C00085), welcher in der VGL1-Modellierung nur 3 anstelle von 17 Reaktionsschritten lang ist. Der Reaktionsweg (R00299) von D-Glucose (C00031) zu D-Glucose 6-phosphate (C00092) ist in diesem Falle der entscheidende Schritt, welcher die Abkürzung ermöglicht. Er wird durch das Glucokinase-Enzym (EC 2.7.1.2) katalysiert, welches schon früh in *C. glutamicum* experimentell nachgewiesen werden konnte (Mori und Shiio, 1987). Es ist sowohl in der Bielefelder- als auch der Kyowa Hakko Annotation vorhanden, fehlt aber als Reaktionseintrag in den entsprechenden Reaktionsnetzwerken. Die Ursache hierzu ist unklar. Durch die Durchführung der Genomannotation im Rahmen dieser Arbeit konnte der Reaktionsweg schließlich Berücksichtigung finden. Sämtliche oben genannten Unterschiede sind in erster Linie auf das erweiterte Enzymrepertoire zurückzuführen, da das KEGG-Mapping in allen Versuchen konstant beibehalten wurde.

7.2.2.3.2 Unter CUBIC-Bedingungen

Analog zum Vorgehen im vorangegangenen Kapitel wurden die Unterschiede der Modellierungen unter CUBIC-Bedingungen systematisch analysiert. In diesem Falle betrug die gemeinsame Schnittmenge zwischen der CGB- und der CGL-Modellierung, wie bereits im Vorfeld erwähnt, 3738 gültige Pfadkombinationen. Unter identischen Bedingungen liefert die Netzwerkmodellierung auf VGL1 insgesamt 4682 gültige Pfade. Die neu hinzugekommenen Pfade teilen sich auch hier auf relativ wenige Metaboliten auf.

So werden bei der Netzwerkmodellierung basierend auf VGL1 unter Verwendung des CUBIC-Mappings nun ebenfalls metabolische Pfade gefunden, welche von alpha,alpha-Trehalose (C01083) ausgehen. Ferner finden sich nun Pfade, die auf D-Fructose 1,6-bisphosphate (C00354), einen Metaboliten der Glykolyse, oder auf LL-2,6-Diamino- heptanedioate (C00666) und L-Lysine (C00047), beides Metaboliten aus der Lysin-Biosynthese, enden. Für die gemeinsame Schnittmenge von 3738 gültigen Pfadkombinationen erbringt die Netzwerkmodellierung auf VGL1 die jeweiligen Entsprechungen zu 100%, das heißt alle „alten“ Pfadkombinationen konnten wiedergefunden werden.

Untersucht man analog auch hier die Pfadlänge der gleichsam ermittelten Pfade, so kommt es erwartungsgemäß wieder zu dem Effekt, dass das CUBIC-Mapping kürzere Pfade liefert als KEGG. Abweichungen in der Pfadlänge treten in 868 der 3738 gemeinsamen Pfade auf, was einem prozentualem Anteil von ca. 23% entspricht. Die Abweichungen sind ausschließlich derart, dass bei der VGL1-Modellierung die kürzeren Pfade gefunden werden können. Die maximale Differenz in der Pfadlänge beträgt unter CUBIC-Bedingungen 7 Reaktionsschritte. So verkürzt sich auch hier der metabolische Pfad von Sucrose (C00089) zu D-Fructose 6-phosphat (C00085) von 10 auf 3 Reaktionsschritte. Der mittlere Unterschied in der Pfadlänge zwischen allen gemeinsamen Paarungen liegt bei 2,36 Schritten.

Die Kombination aus der Anwendung des CUBIC-Mappings und der Verwendung des VGL1-Netzwerkes führt zu teilweise tiefgreifenden Veränderungen in der Modellierung der Stoffwechselwege. Die Abbildung 7.11 erlaubt einen Vergleich darüber, wie sich die Pfadlänge in Abhängigkeit der Reaktionsnetzwerke und beider Mapping-Algorithmen verhält. Datengrundlage ist, wie bereits angesprochen, die Schnittmenge zwischen der CGB- und der CGL-Modellierung, sowie deren Entsprechung in der VGL1-Modellierung. Wären sämtliche Pfade gleich lang, lägen alle Datenpunkte auf einer Geraden, was nicht der Fall ist. Man kann er-

kennen, dass die Unterschiede in der Pfadlänge unter KEGG-Bedingungen (blaue Punkte) größer sind als unter CUBIC-Bedingungen (rote Punkte). Auffällig ist auch, dass eine Reihe von Pfadkombinationen, deren Länge in der CGB- und CGL-Modellierung zwischen 12 und 16 Reaktionsschritte betrugen, durch die Erweiterung des enzymatischen Repertoires deutlich gekürzt werden konnten.

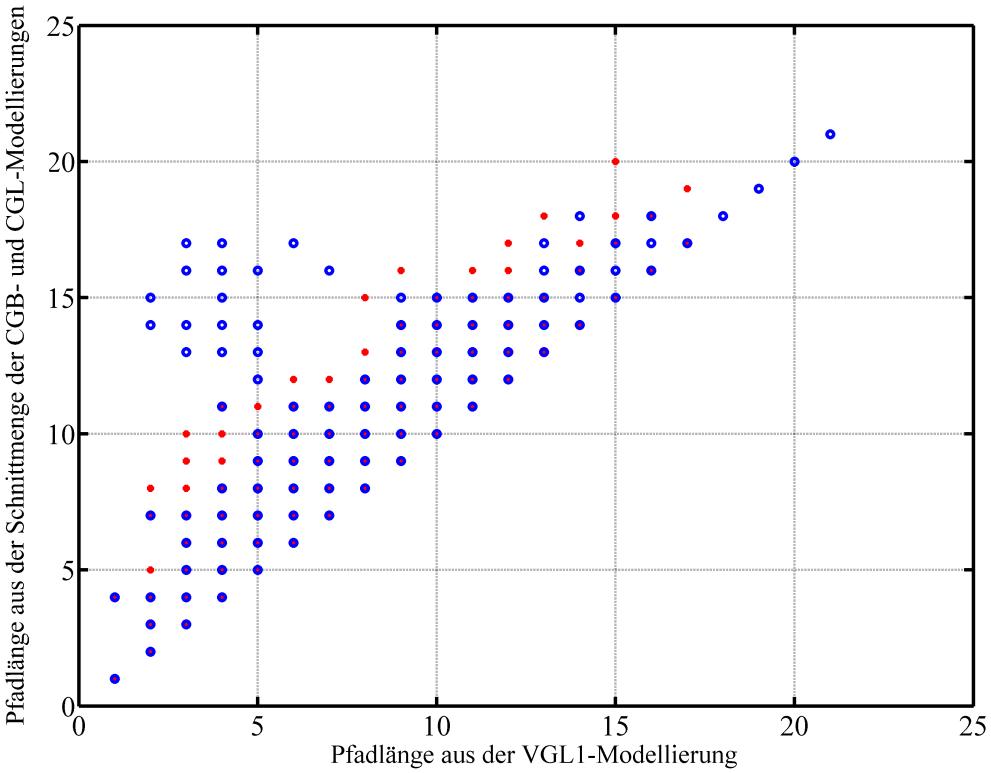

Abbildung 7.11: Vergleich der Pfadlängen der gleichermaßen in der CGB-, CGL- sowie VGL1-Modellierung gefundenen metabolischen Pfade. Betrachtung unter KEGG- (blau) und CUBIC-Bedingungen (rot). Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit wurden verwendet. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen.

7.2.3 Zusammenfassende Betrachtung

Die primäre Betrachtung der verschiedenen Modellierungen der Stoffwechselwege förderte interessante Erkenntnisse zu Tage. Es kann gesagt werden, dass sowohl die Verwendung unterschiedlicher Annotationen für *C. glutamicum* als auch die Wahl der Mapping-Algorithmen zu unterschiedlichen Ergebnissen führt. Generell konnte bei allen Modellierungen festgestellt werden, dass die Verwendung des CUBIC-Mapping mehr und im Mittel kürzere Pfade erbringt, als die Verwendung des KEGG-Mappings. Die alleinige Verwendung des CUBIC-Algorithmus bei sonst konstanten Bedingungen führte bei der CGB- und CGL-Modellierung zu einer Steigerung der Anzahl gefundener Pfade von über 50%. Bei den eigenen Modellierungen (VGL1 und VGL2) betrug die Steigerung hingegen nur 25%. Die Ursache liegt darin begründet, dass beim CUBIC-Mapping generell eine höhere Anzahl von Verknüpfungsmöglichkeiten zwischen Metaboliten zugelassen wird, was vor allem bei schwach konnektierten Netzwerken größere Effekte zeigt.

Die Verwendung der erweiterten Reaktionsnetzwerke VGL1 und VGL2 führte ebenfalls zu einer deutlichen Steigerung der Anzahl gefundener Pfade. Da diese Steigerung als direkte Konsequenz des erweiterten enzymatischen Repertoires zu sehen ist und im Gegensatz zur Verwendung des CUBIC-Mappings nachvollziehbare Veränderungen zeigte, wurde das KEGG-Mapping im weiteren Verlauf der Arbeit dem CUBIC-Mapping vorgezogen.

Generell konnte beim Vergleich zur neuen VGL1-Modellierung festgestellt werden, dass ein Zuwachs von rund 10% mehr Enzymen unter KEGG-Bedingungen in rund 46% mehr Pfaden und unter CUBIC-Bedingungen in 20% mehr Pfaden resultiert. Im Umkehrschluss kann angemerkt werden, dass sich das KEGG-Mapping auf die Vergrößerung des Enzymrepertoires deutlich sensitiver in einer Erhöhung der gefundenen Pfade reagiert als das CUBIC-Mapping, dessen Einfluss - wie bereits erwähnt - nicht reproduzierbar nachvollzogen werden kann. Betrachtet man die Auswirkung des Mapping-Algorithmus ferner hinsichtlich der Länge der gefundenen Pfade, so hat auch hier unter KEGG- Bedingungen die Vergrößerung des Repertoires von Enzymen einen größeren Effekt als unter CUBIC-Bedingungen.

7.3 Analyse der abgeleiteten Deskriptorensets

Auf eine detaillierte Analyse der abgeleiteten experimentellen und theoretischen Deskriptoren untereinander wurde an dieser Stelle zugunsten der gemeinsamen, integrativen Analyse beider Datensätze verzichtet. An dieser Stelle soll lediglich auf einige auffällige Zusammenhänge zwischen Deskriptoren hingewiesen werden.

7.3.1 Experimentelle Deskriptoren

Die für den paarweisen Vergleich zweier Metabolitzeitreihen berechneten Deskriptoren weisen oft eine Ähnlichkeit zueinander auf. So ist beispielsweise in der Abbildung 7.12 gut zu erkennen, dass zwischen dem Korrelationsmaß und der Winkelähnlichkeit sowie der Gleichläufigkeit ein Zusammenhang existiert.

Das zur Beschreibung der Formähnlichkeit zweier Zeitreihen berechnete Winkelmaß verhält sich zur Korrelation gegenläufig, das heisst, den höchsten Korrelationswerten entspricht der niedrigste Wert der Winkelähnlichkeit. Mit absteigender Korrelation fasert das Winkelmaß stärker aus. Da nur wenige antikorrelierte Paarungen vorliegen, kann keine Aussage für den Bereich der stark negativen Korrelationen getroffen werden. Zu vermuten ist, dass das Winkelmaß in seiner Streuung wieder abnimmt. Die Gleichläufigkeit, welche ebenfalls die Formähnlichkeit der Zeitreihen zueinander beschreibt, deckt - wie bereits in Kapitel 5.1.3.3 beschrieben - keinen kontinuierlichen Wertebereich ab. Dennoch kann auch hier ein, wenngleich wenig deutlicher Zusammenhang zur Korrelation festgestellt werden. Die Streuung der Gleichläufigkeit ist in allen Wertebereichen der Korrelation annähernd gleich groß.

7.3.2 Theoretische Deskriptoren

Die Betrachtung der theoretischen Deskriptoren ist im Gegensatz zu den experimentellen Deskriptoren deutlich heterogener. Es zeigt sich, dass zwar keine deutlichen Abhängigkeiten, jedoch Auffälligkeiten zwischen Deskriptoren existieren.

So weisen zum Beispiel die Länge eines gefundenen metabolischen Pfades und die Anzahl alternativer, gleich langer Pfade eine Auffälligkeit auf. Abbildung 7.13 zeigt exemplarisch diesen Zusammenhang bei der CGB-Modellierung unter KEGG-Bedingungen. Im Durchschnitt existieren für jeden Pfad rund 1,5 alternative Pfade. Wie die Grafik zeigt, werden bei sehr kurzen Pfaden in der Regel

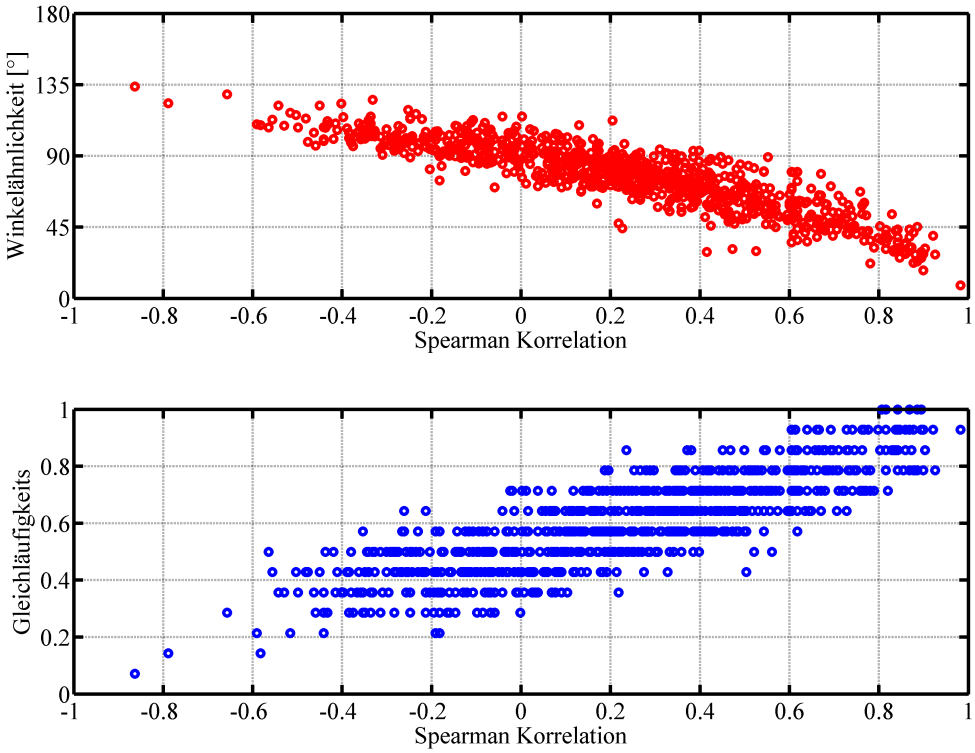

Abbildung 7.12: Zusammenhang zwischen der Spearman'schen Korrelation und dem Winkelmaß (rot), sowie zwischen Spearman'schen Korrelation und der Gleichläufigkeit (blau). Basierend auf exemplarischen Daten der Glucose-Fermentation unter Betrachtung der exponentiellen Wachstumsphase. Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung.

nur wenige Alternativen gefunden. Die Anzahl der kürzesten Pfade steigt mit zunehmender Pfadlänge an und erreicht zwischen 10 und 13 Reaktionsschritten ihr Maximum, bei größeren Pfadlängen nimmt die Anzahl verfügbarer Alternativpfade erneut ab. Diese Auffälligkeit könnte mit der Struktur des Netzwerkes zusammenhängen. Extrem lange Pfade existieren beispielsweise von Peripherie zu Peripherie, wobei der Zentralstoffwechsel durchquert wird. Es ist anzunehmen, dass für diesen seltenen Spezialfall wenn überhaupt nur wenige Alternativen existieren können. Die Tatsache, dass das Maximum verfügbarer Alternativen ca. zwischen 10 und 13 Reaktionsschritten liegt, könnte ein Indikator dafür sein, dass sich der Netzwerkradius in diesen Größenordnungen bewegt. Aufgrund der in Kapitel 7.2.2 durchgeführten Versuche ist ferner davon auszugehen, dass die Abhängigkeit zwischen der Pfadlänge und dem Vorhandensein alternativer, gleich langer Pfade stark vom Mapping-Verfahren sowie dem zugrunde liegenden Reaktionsnetzwerk abhängig ist.

7.3.3 Zusammenfassende Betrachtung

Die Tatsache, dass experimentelle Deskriptoren teilweise deutliche Abhängigkeiten zueinander aufweisen, ist nachvollziehbar, schließlich wurden sie auf den identischen Ausgangsdaten zum Zwecke der Ähnlichkeitsbeschreibung berechnet. Anzumerken bleibt jedoch, dass Deskriptoren alles andere als deckungsgleich sind, da sie unterschiedliche Eigenschaften aus den Zeitreihen abgreifen. Die Untersuchung der theoretischen Deskriptoren kann erste Aussagen über die zugrundeliegende Struktur des metabolischen Netzwerkes erbringen und sollte daher weitergehend untersucht werden.

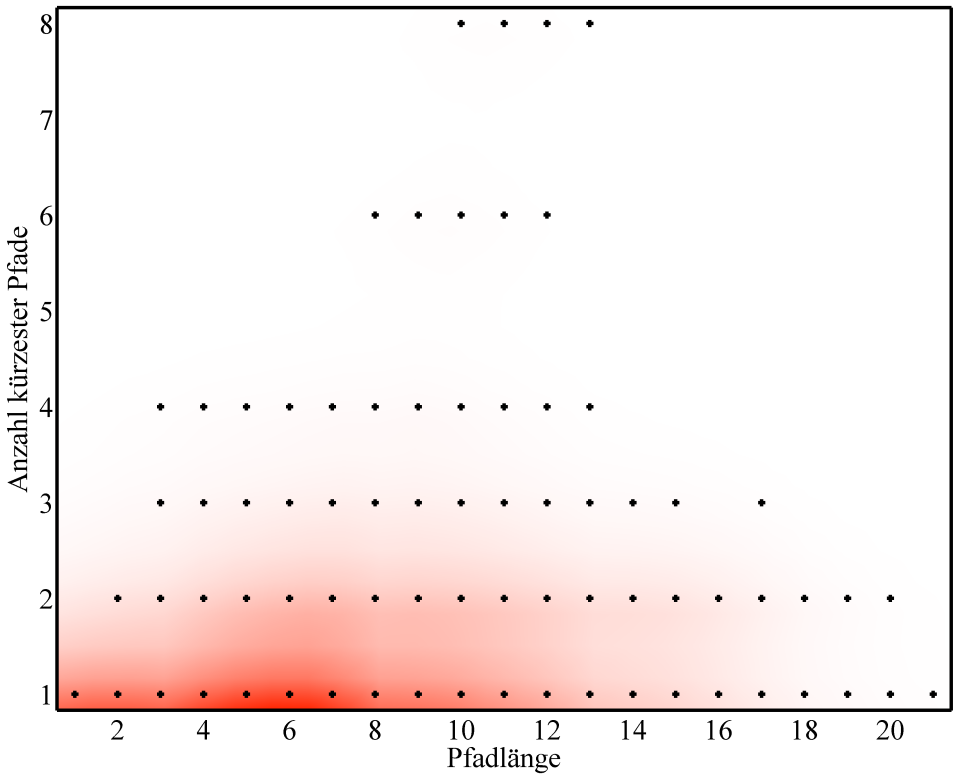

Abbildung 7.13: Zusammenhang zwischen der Pfadlänge und der Anzahl gefundener Pfade am Beispiel der CGB-Modellierung unter KEGG-Bedingungen. Die Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit wurden verwendet. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen. Anzahl der Datenpunkte: 2559. Interpolierte Dichtefunktion.

7.4 Integrative Analyse experimenteller und theoretischer Deskriptoren

In diesem Teil der Arbeit wird die Schnittmenge der experimentellen und theoretischen Daten in detaillierter Form betrachtet. Dies bedeutet, dass nur Metaboliten betrachtet werden, für die sowohl experimentelle Informationen als auch Informationen aus der theoretischen Betrachtung vorliegen. Es werden gleichermaßen die Eigenschaften einzelner Metaboliten als solche, aber auch deren paarweise Beziehungen zueinander untersucht. Die beschreibenden Attribute stehen, wie im Vorfeld erläutert, auf experimenteller und theoretischer Seite in Form von abgeleiteten Deskriptoren zur Verfügung. Ziel der integrativen Analyse war es, festzustellen, ob sich Eigenschaften der experimentell erfassten Metabolitzeitreihen durch die Zusammenschau mit den theoretischen metabolischen Informationen erklären lassen.

Die modulare Konzeption der Arbeit erlaubt es, den systematischen Einfluss von Effekten auf die Analyse und deren Ergebnisse zu untersuchen. Was die experimentellen Daten betrifft, so sind dies in erster Linie die fünf verschiedenen Fermentationsexperimente, in denen der Stoffwechsel - wie in den Zeitreihen der Metabolitkonzentration sichtbar - unterschiedlichen Gesetzmäßigkeiten folgt. Die unterschiedlichen Ansätze der Datenvorverarbeitung wurden ebenfalls als induzierte Effekte untersucht. Auf der Seite der theoretischen Betrachtung lagen Deskriptorensätze vor, welche aus den jeweiligen Netzwerkmodellierungen abgeleitet wurden. Da auch die Modellierungen auf verschiedenen Reaktionsnetzwerken und unter Verwendung verschiedener Einstellungen durchgeführt worden sind, bietet sich auch hier die systematische Untersuchung dieser Effekte an.

Wie aufgrund der hohen Anzahl induzierter Effekte deutlich wird, ergeben sich für die Betrachtung von Zusammenhängen zwischen experimentellen und theoretischen Deskriptoren sehr viele mögliche Kombinationen. Aus diesem Grunde wurde die Untersuchung nach vorhandenen Auffälligkeiten in Form von automatisierten Batch-Prozessen durchgeführt.

Die Durchsicht der Ergebnisse der Batch-Prozesse ergab, dass generell nur für sehr wenige Kombinationen experimenteller und theoretischer Deskriptoren tatsächlich ein Zusammenhang detektiert werden konnte. Die Gründe hierfür, sowie die wichtigsten Erkenntnisse der integrativen Analyse werden in den nachfolgen-

den Kapiteln detailliert beschrieben. Zuerst werden Auffälligkeiten bei der integrativen metabolit-zentischen und anschließend bei der integrativen paarweisen Betrachtung von Metaboliten und deren theoretischen Eigenschaften dargelegt.

7.4.1 Integrative Analyse metabolitspezifischer Merkmale

Hierunter ist zu verstehen, dass der einzelne Metabolit und seine Eigenschaften als solche im Fokus der Betrachtung stehen. Oder vereinfacht ausgedrückt, es wurden gemessene Eigenschaften eines Metaboliten den theoretischen Eigenschaften desselben gegenübergestellt. Dementsprechend wurden Zeitreiheneigenschaften einzelner Metaboliten mit theoretischen Deskriptoren verglichen. Interessante Ergebnisse konnten bei einigen Parameterkombinationen festgestellt werden, die nachfolgend wiedergegeben sind. Insbesondere wurde deutlich, dass die Position eines Metaboliten in seinem Netzwerk - beschrieben durch seinen Verknüpfungsgrad - wertvolle Informationen liefert.

7.4.1.1 Metabolitkonzentration gegen Verknüpfungsgrad aus KEGG-Datenbank

In einem ersten Schritt wurde untersucht, ob ein Zusammenhang zwischen der Konzentration eines Metaboliten und seinem theoretischen Verknüpfungsgrad (Linkage) zu anderen Metaboliten existiert. Bei den experimentellen Daten dieser Arbeit handelt es sich, wie in Kapitel 4.2.3 dargelegt, um semi-quantitative Daten. Die metabolitspezifischen Faktoren zur Umrechnung der Peakflächen in tatsächliche Konzentrationswerte lagen zum Zeitpunkt der Durchführung dieser Arbeit noch nicht vor. Obwohl in dieser Arbeit die Zeitreiheneigenschaften von Metaboliten und weniger deren absolute Konzentration im Vordergrund des Interesses stehen, soll an dieser Stelle die Konzentration als beschreibende Größe herangezogen werden. Zur Skalierung der Daten wurden die logarithmierten Daten nach adaptiver Fehlerkorrektur und Ausreißerkorrektur untersucht.

Der Verknüpfungsgrad als theoretische Größe ist hierbei durch die Anzahl von Enzymen, welche den Metaboliten umwandeln können, gekennzeichnet. Die Verknüpfungsinformation wurde (ähnlich wie in Kapitel 4.3.3 zur Defintion von Seitenmetaboliten) aus der KEGG-Datenbank entnommen. Verwendet wurde anstelle des generellen, organismenübergreifenden Referenzstoffwechsels an dieser Stelle jedoch organismenspezifische Information für *C. glutamicum*. Die Abbildung 7.14

verdeutlicht den Sachverhalt am Beispiel der Glucose-Fermentation. Dargestellt ist, wie sich die Konzentrationsverläufe der Metaboliten zum Verknüpfungsgrad verhalten.

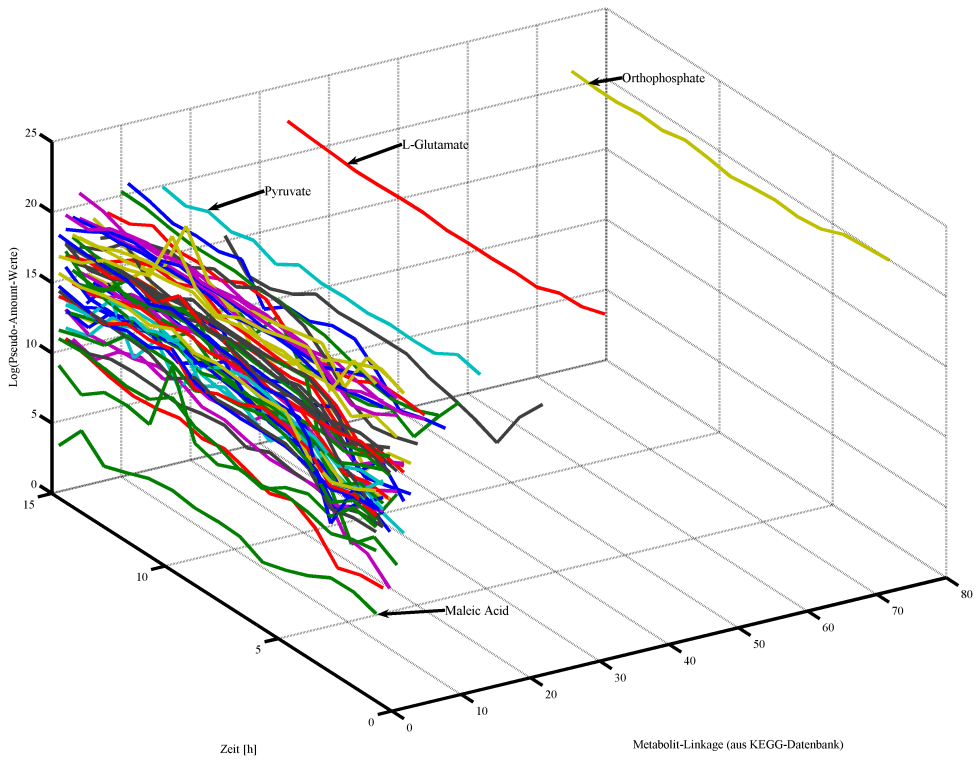

Abbildung 7.14: Zusammenhang zwischen der Metabolitkonzentration und dem aus der KEGG-Datenbank abgeleiteten Verknüpfungsgrad. Datengrundlage: logarithmierte Daten der Glucose-Fermentation nach adaptiver Fehlerkorrektur und Ausreißerkorrektur.

Es zeigte sich, dass in sehr geringen Konzentrationen auftretende Metaboliten wie Maleic Acid (C01384) eine sehr geringe Verknüpfungszahl im theoretischen metabolischen Netzwerk aufweisen. Umgekehrt ist zu erkennen, dass die höchsten Verknüpfungsgrade ausschließlich bei jenen Metaboliten gefunden wurden, die in vergleichsweise hohen Konzentrationen vorkommen (z.B. Orthophosphate). Stark verknüpfte Metaboliten, die gleichzeitig eine niedrige Konzentration aufweisen, konnten nicht festgestellt werden. Zusammengefasst kann gesagt werden, dass in geringen Konzentrationen auftretende Metaboliten tendenziell eher eine niedrige

Verknüpfungszahl aufweisen, also unter Umständen eher an der Peripherie des metabolischen Netzwerkes anzusiedeln sind. Diese Auffälligkeit konnte auch bei den Fermentationen mit Acetat, Fructose, Glutamin und Lactat festgestellt werden, wenngleich nicht in der Deutlichkeit der Glucose-Fermentation.

Eine mögliche Erklärung für dieses Phänomen könnte sein, dass beispielsweise ein Metabolit wie Orthophosphate (C00009), welcher bei der Suche von metabolischen Pfaden in theoretischen Netzwerken als Seitenmetabolit deklariert wurde, in vergleichsweise hohen Konzentrationen existieren muss, da er an vielen essentiellen Reaktionen beteiligt ist. Für die Metaboliten L-Glutamate (C00025) und Pyruvate (C00022) trifft dies nicht zu. Hier handelt es sich um Beispiele der bereits in der theoretischen Betrachtung metabolischer Netze angesprochenen Metabolit-Hubs, welche für die Synthese zahlreicher anderer Metaboliten und damit für das Überleben des Organismus essentiell sind. Ihr Vorkommen in vergleichbar hohen Konzentrationen kann dadurch erklärt werden, dass sie Bestandteile vieler Flüsse sind, die zudem in vergleichbar großen Flussraten existieren.

Ein Metabolit, welcher nur in sehr geringen Konzentrationen gemessen werden konnte ist Maleic Acid (C01384). Interessanterweise ist Maleic Acid in allen Modellierungen nur durch eine einzige Reaktion (R03540) mit einem anderen Metaboliten verbunden, wobei es sich um Maleamat (C01596) handelt. Beide Metaboliten bilden also nach dem theoretischen Wissenstand für *C. glutamicum* ein isoliertes Subnetz im metabolischen Netzwerk. Schaut man in den, in der KEGG-Datenbank hinterlegten organismenübergreifenden Referenzstoffwechsel, so existieren überhaupt nur drei biochemische Reaktionen, die diesen Metaboliten synthetisieren können. Zwei dieser Alternativen sind relativ unspezifisch und nicht durch Sequenzinformation belegt, eine scheint jedoch im Kontext dieser Arbeit wahrscheinlicher. Hierbei handelt es sich um die Verknüpfung von Maleic Acid zu Fumarate (C00122), wie sie nachweisbar in anderen Bakterien (beispielsweise dem ebenfalls im Boden vorkommendem Bakterium *Serratia marcescens*) nachweislich existiert (Hatakeyma et al., 2000). Bei dem betreffenden Enzym handelt es sich um das Enzym Maleate Isomerase (EC 5.2.1.1), welche die entsprechende Reaktion (R01087) katalysiert. Theoretisch angenommen, dass diese Reaktion das bisherige Subnetzwerk zum restlichen Stoffwechsel konnektiert, könnte Maleic Acid nur aus Fumarate synthetisiert werden. In diesem Falle würde seine Konzentration gänzlich von der Konzentration von Fumarate und der enzymatischen Aktivität des katalysierenden Enzyms abhängig sein. Aufgrund dieser Abhängigkeit, ist

wahrscheinlich die Konzentration, in der Maleic Acid überhaupt auftreten kann, stark limitiert. Nicht zuletzt dadurch, dass Fumarate als Metabolit des Zitratzyklus mit Sicherheit deutlich intensiver mit seinen Nachbarn (S)-Malate (C00149) und Succinate (C00042) interagiert. Es ist anzunehmen, dass die Intensität des metabolischen Flusses zu Maleic Acid nur relativ gering sein kann.

7.4.1.2 Metabolitkonzentration gegen Verknüpfungsgrad aus PHT-Modellierung

Alternativ zum Verknüpfungsgrad aus der KEGG-Datenbank, kann die Verknüpfung eines Metaboliten auch aus der Modellierung mit dem Pathway Hunter Tool abgeleitet werden. Vergleicht man nun die Konzentrationswerte mit dem Verknüpfungsgrad aus der Modellierung in Abbildung 7.15, so ergibt sich ein ähnliches Bild. Die errechneten Verknüpfungsgrade aus den Modellierungen ähneln tendenziell denen der KEGG-Datenbank. Maleic Acid (C01384), welches - wie bereits erwähnt - nur in geringen Konzentrationen erfasst worden ist, ist auch hier der Metabolit mit der niedrigsten Verknüpfungszahl, während Orthophosphate (C00009), den höchsten Verknüpfungsgrad besitzt. Ein annähernd gleich hoher Verknüpfungsgrad für die Metaboliten L-Glutamate und Pyruvate ist hierbei nicht festzustellen. Erneut zu erkennen ist allerdings, dass Metaboliten hoher Konzentration nie bei niedrigen Verknüpfungsgraden auftauchen. Die Tatsache, dass Orthophosphate in dieser Liste vorkommen, obwohl er explizit als Seitenmetabolit bei der Pfadsuche deklariert wurde ist einfach zu erklären. Die Berechnung der Verknüpfungen (Linkages) also auch der Ladungen (Loadings), vergleiche hierzu im Detail die Kapitel 5.2.4.6 und 5.2.4.8 wird unabhängig von der Pfadsuche durchgeführt. Bei der eigentlichen Pfadsuche werden deklarierte Seitenmetaboliten - wie beabsichtigt - ausgeschlossen.

7.4.1.3 Metabolitkonzentration gegen Anzahl hindurchgehender Pfade

Das Pathway Hunter Tool erlaubt es auch, zu berechnen, wie viele metabolische Pfade unter den gegebenen Randbedingungen ihren Weg über einen bestimmten Metaboliten innerhalb des Netzwerkes nehmen (vergleiche hierzu Kapitel 5.2.4.7). Untersucht man den Zusammenhang zur Konzentration des entsprechenden Metaboliten, so ergibt sich ein Sachverhalt, wie er in Abbildung 7.16 dargestellt ist. Es zeigt sich, dass generell ein wenig deutlicher Zusammenhang festgestellt

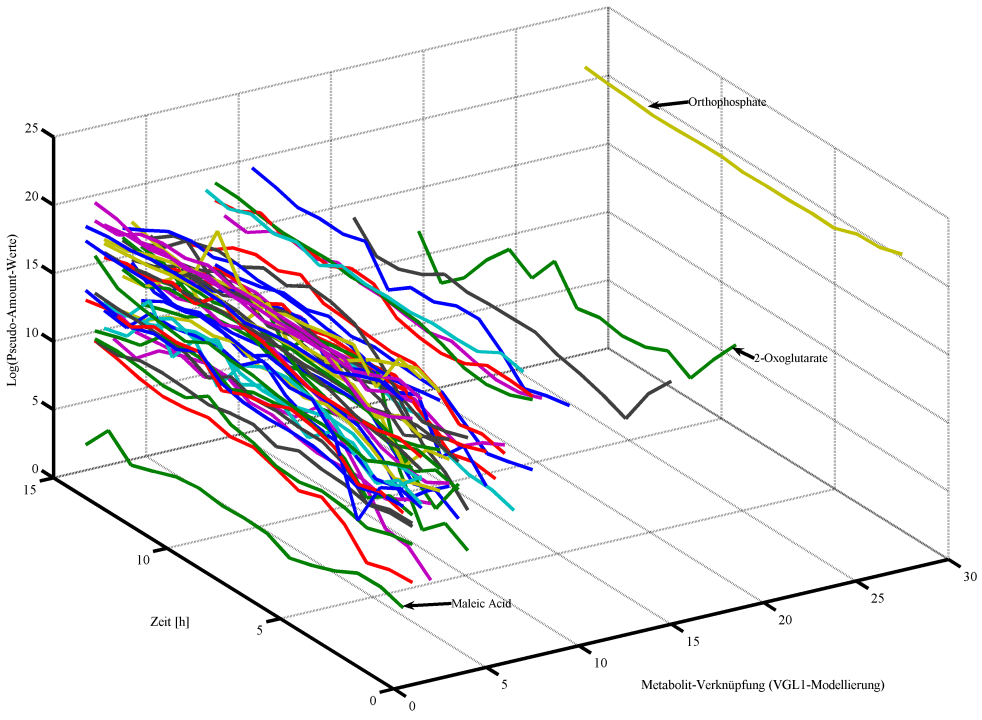

Abbildung 7.15: Zusammenhang zwischen der Metabolitkonzentration un dem aus der PHT-Analyse abgeleiteten Verknüpfungsgrad. Theoretische Parameter: Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit, sowie Verwendung des KEGG-Mapping auf dem VGL1-Netzwerk. Experimentelle Datengrundlage: logarithmierte Daten der Glucose-Fermentation nach adaptiver Fehlerkorrektur und Ausreißerkorrektur.

werden kann. So weisen beispielsweise die Metaboliten 2-Oxoglutarate (C00026) und D-Ribose 5-Phosphate (C00117), die essentielle Funktionen im Zitratzyklus, respektive im Pentose-Phosphat-Weg innehaben, bei mittelwertigen Konzentrationen eine sehr hohe Anzahl von theoretisch durch sie hindurchgehenden Pfaden auf. Metaboliten, über die eine hohe Anzahl von (theoretischen) metabolischen

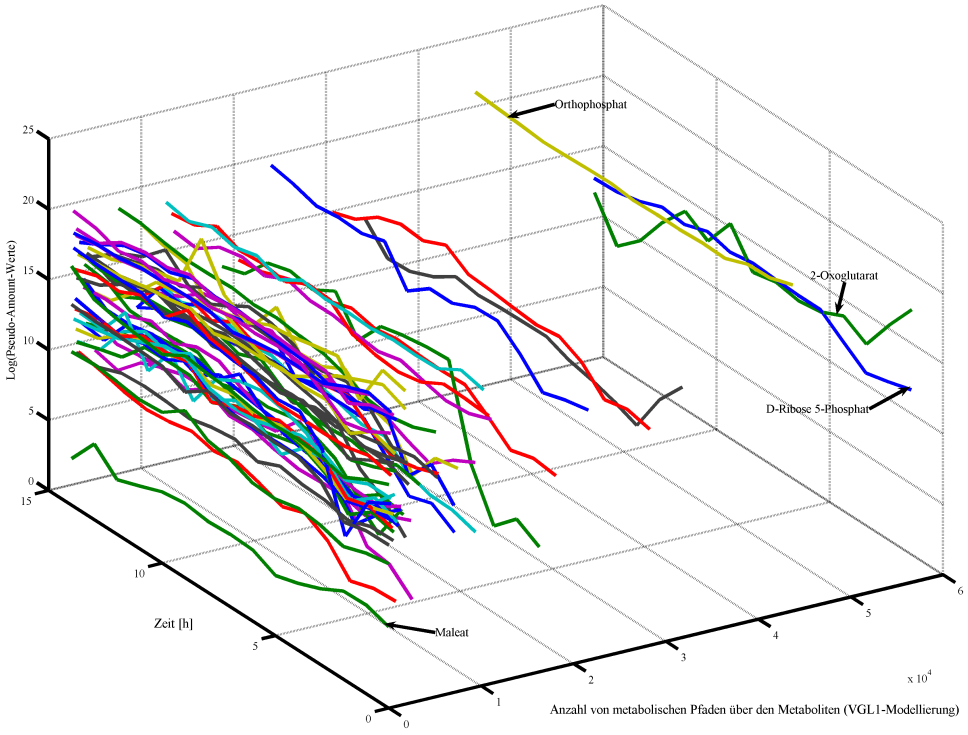

Abbildung 7.16: Zusammenhang zwischen der Metabolitkonzentration und der Anzahl metabolischer Pfade, die über den betreffenden Metaboliten führen. Theoretische Parameter: Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit, sowie Verwendung des KEGG-Mapping auf Reaktionsnetzwerk VGL1. Experimentelle Datengrundlage: logarithmierte, ausreißer- und nullwertkorrigierte Daten der Glucose-Fermentation

Pfaden führt, sind nicht bei größten Konzentrationen vorhanden, sondern weisen Konzentrationen auf, die im mittleren Bereich liegen. Dies weist darauf hin, dass nicht zwangsläufig eine hohe Konzentration ein Maß für die Wichtigkeit eines Me-

taboliten im Stoffwechsel ist. Diese Erkenntnis steht in Übereinstimmung mit den Arbeiten von van den Berg et al., 2006. Beachtet werden muss allerdings, dass die Anzahl durch einen Metaboliten hindurchgehender Pfade ein theoretischer Deskriptor ist, welcher sehr stark von den gegeben Startbedingungen des Pathway Hunter Tool anhängig ist. Die Untersuchung der Metabolitladungen (Loadings, vergleiche hierzu insbesondere Kapitel 5.2.4.8) in Zusammenhang mit den Konzentrationsdaten erbrachte keinerlei Auffälligkeiten. Vermutlich deshalb, da die Metabolit-Ladungen als solche noch stärker von den Startbedingungen der Modellierung des PHT abhängig sind, als beispielsweise der Verknüpfungsgrad.

Für die beobachteten Fermentationen kann zusammenfassend gesagt werden, dass die Konzentration in der ein Metabolit detektiert wird (eingeschränkte) Rückschlüsse darüber zulässt, in welcher Position beziehungsweise Funktion sich ein Metabolit innerhalb des Netzwerkes befindet. Niedrige Konzentrationen besitzen tendenziell eher jene Metaboliten, die ihrerseits wenige Nachbarn besitzen und unter Umständen an der Peripherie des Netzwerkes anzusiedeln sind. Jene Metaboliten, welche einen ausgesprochen hohen Verknüpfungsgrad aufweisen, treten in vergleichsweise höheren Konzentrationen auf. Die statischen Verknüpfungsinformation aus KEGG, sowie die aus der Modellierung mit dem Pathway Hunter Tool abgeleiteten Informationen weisen gegenseitige Ähnlichkeiten auf.

7.4.1.4 Sensitivität der Zeitreihe gegen Verknüpfungsgrad aus KEGG-Datenbank

Bringt man anstelle der Konzentration eine andere Eigenschaft der Zeitreihe - wie beispielsweise die Schwankungsbreite - ins Spiel, so ergibt sich in Zusammenschau mit dem Verknüpfungsgrad ein äußerst interessantes Bild. Abbildung 7.17 zeigt, dass Metaboliten, die eine hohe Schwankungsbreite aufweisen, tendenziell wenige Nachbarn besitzen. Umgekehrt fällt auf, dass stark verknüpfte Metaboliten tendenziell eher eine geringere Schwankungsbreite aufweisen. Hohe Sensitivitäten von Zeitreihen bei gleichzeitig hohem Verknüpfungsgrad sind nicht festzustellen.

Es ist deutlich festzustellen, dass die höchsten Schwankungsbreiten bei jenen Metaboliten festzustellen sind, die wenige Nachbarn besitzen. Je stärker ein Metabolit verknüpft ist, desto häufiger tritt er als Reaktionspartner in Aktion. Es ist möglich, dass sich im Mittel die Reaktionen in ihrer Wirkung ausgleichen und dies zu vergleichsweise konstanten Verläufen der Konzentration führt. Hinzu kommt,

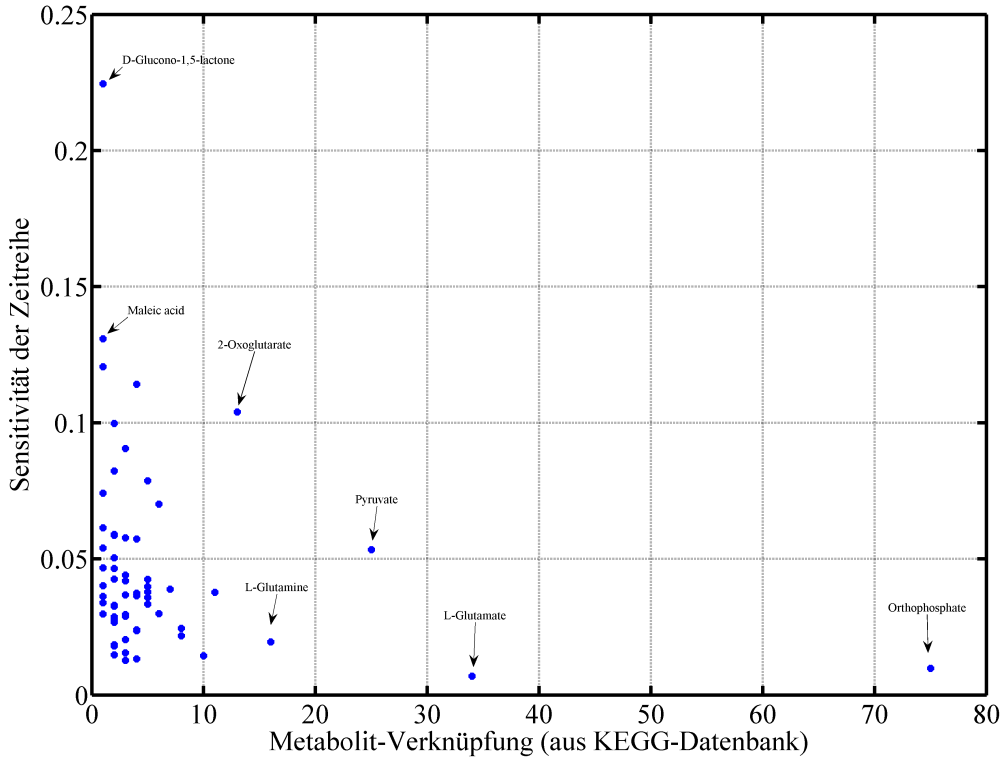

Abbildung 7.17: Zusammenhang zwischen der Sensitivität der Konzentrationszeitreihe und dem Verknüpfunsgrad aus KEGG-Datenbank. Experimentelle Datengrundlage: ausreißer-und nullwertkorrigierte, logarithmierte und medianzentrierte Daten der Glucose-Fermentation.

dass stark verknüpfte Metaboliten, wie bereits gezeigt, in vergleichsweise hohen Konzentrationen auftreten. Unter Umständen sind diese „Pools“ so groß, dass sich der Einfluss anderer Reaktionen kaum bemerkbar macht. Diese Erkenntnis könnte unter Umständen dazu benutzt werden, den theoretischen Verknüpfungsgrad eines Metaboliten aus der Sensitivität seiner Zeitreihe abzuschätzen. Dieses Vorhaben setzt allerdings voraus, dass der Versuchsaufbau, sowie die Datenprozessierung zu standardisieren sind.

7.4.2 Integrative Analyse paarweiser Metaboliteigenschaften

Hierbei steht das gegenseitige Verhältnis zweier Metaboliten zueinander im Mittelpunkt des Interesses. Dies bedeutet, dass sowohl auf der experimentellen als auch theoretischen Seite Informationen verwendet werden, welche die Beziehung zweier Metaboliten zueinander charakterisieren. Hierzu finden die in Kapitel 5.1.3 entwickelten Deskriptoren Anwendung.

Wie auch in der metabolitzentrischen Betrachtung, konnten für die paarweise Analyse in der Mehrzahl der betrachteten Deskriptorenkombinationen keine signifikanten Zusammenhänge in der automatisierten Suche detektiert werden. Es zeigte sich, das die Ursachen hierfür in detaillierten Untersuchungen einzelner Kombinationen von experimentellen und theoretischen Deskriptoren zu beleuchten sind. Diese Vorgehensweise ist in den nachfolgenden Unterkapiteln dargestellt.

7.4.2.1 Prozessähnlichkeit und theoretischer Reaktionsabstand

In zahlreichen Untersuchungen (Kose et al., 2001; Steuer et al., 2003 oder Weckwerth et al., 2004) wurde die paarweise Korrelation zwischen Metaboliten (auf der Basis punktueller Konzentrationsmessungen und zur Verfügung stehenden Replikaten) untersucht. Die Ergebnisse zeigten, dass deutliche Korrelationen zwischen Metaboliten existieren und das jene das Ergebnis enzymatisch regulierter Zusammenhänge sein können.

Aufgrund dieser Erkenntnisse und dem in Kapitel 7.1.4.2 gewonnenen Ergebnis, dass sich die beobachteten Zeitreihen der Metabolitkonzentration deutlich in Gruppen ähnlichen temporalen Verhaltens einordnen lassen, wurde untersucht, ob Ähnlichkeiten zwischen Zeitreihen von dem zugrunde liegenden theoretischen

metabolischen Netzwerk und seinen Eigenschaften abhängig sind. Insbesondere wurde hierbei die paarweise Prozessähnlichkeit in Zusammenschau mit dem korrespondierendem Reaktionsabstand zweier Metaboliten detailliert untersucht. Diese Untersuchung wurde systematisch im Batchverfahren auf allen verfügbaren Netzwerkmodellierungen, die sich (wie in Kapitel 7.2 ausführlich beschrieben), stark unterscheiden können, sowie auf allen zur Verfügung stehenden Fermentationsexperimenten und Datenvorverarbeitungen durchgeführt. Im Gegensatz zu Kapitel 7.1.2 wurde für die Berechnung der paarweisen Prozessähnlichkeit nicht die gesamte zur Verfügung stehende Zeitreihe genutzt, sondern primär der Bereich der exponentiellen Wachstumsphase genutzt. So ist sichergestellt, dass die Daten einer homogenen Grundgesamtheit entstammen.

Die Abbildung 7.18 stellt den Zusammenhang zwischen dem Spearman'schen Korrelationskoeffizienten (paarweise zwischen zwei Metabolitzeitreihen berechnet) und dem korrespondierendem Reaktionsabstand der beiden Metaboliten im metabolischen Netzwerk dar. Folgende erste Aussagen lassen sich aus der Betrachtung der Grafik ableiten:

- Die Mehrzahl der Metabolitpaarungen weist keine oder allenfalls eine moderate Korrelation auf. Dieser Sachverhalt ist in Übereinstimmung mit den Ergebnissen von Roessner et al. (2001) und Weckwerth et al. (2004).

- Negative Korrelationen sind generell weniger häufig. Deutlich antikorrelierte Metabolitzeitreihen sind extrem selten zu finden.

- Deutlich positiv korrelierte Paarungen finden sich in direkter Nachbarschaft als auch in großen Reaktionsabständen zueinander. Dies ist in Übereinstimmung mit den Untersuchungen von Steuer et al. (2003) und Camacho et al. (2005), welche jedoch die Korrelation aufgrund punktueller Konzentrationsmessungen in mehreren Replikaten untersucht haben.

- In direkter Nachbarschaft ist die paarweise Ähnlichkeit der Konzentrationszeitreihen der betreffenden Metaboliten nicht notwendigerweise höher als unter größeren Reaktionsabständen. So konnte für die Spearman'sche Korrelation bei der Untersuchung von Metaboliten in direkter theoretischer Nachbarschaft ein Wertebereich zwischen 1 und 0,4 festgestellt werden. Es ist allerdings auffällig, dass signifikante Antikorrelationen nicht in direkter

Nachbarschaft gefunden werden konnten, sondern - wenn überhaupt - in höheren Reaktionsabständen zu finden sind.

- Bei allen Fermentationen konnte festgestellt werden, dass kein genereller, übergeordneter Zusammenhang zwischen der paarweisen Prozessähnlichkeit (beispielsweise abgegriffen durch die Korrelation) und dem theoretischen Reaktionsabstand zweier Metaboliten nachgewiesen werden kann.

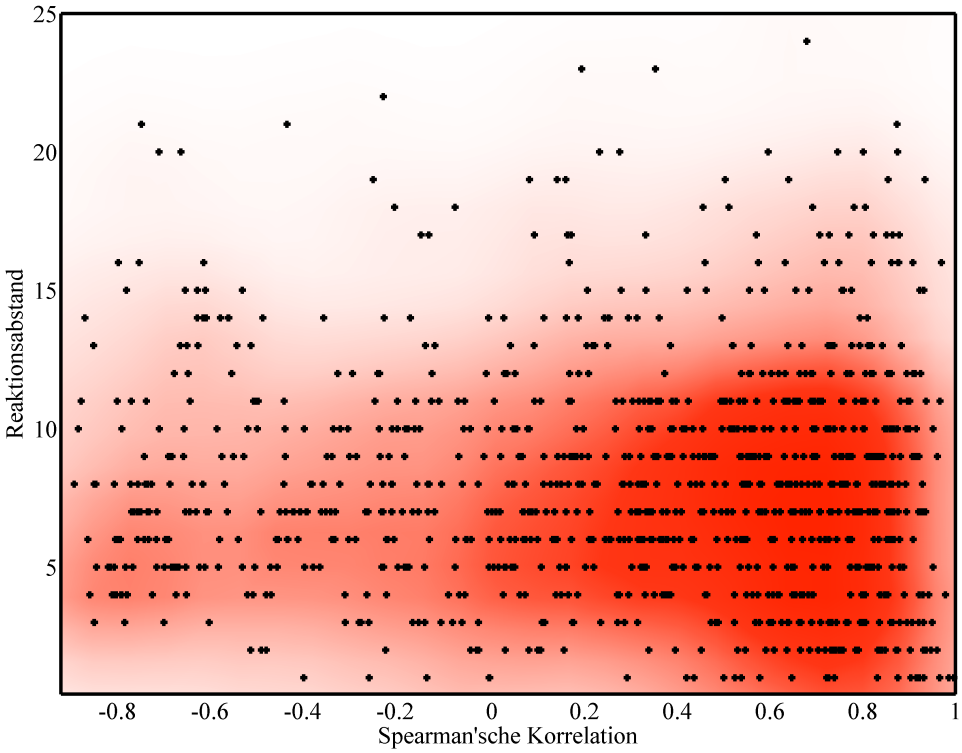

Abbildung 7.18: Zusammenhang zwischen der paarweisen Metabolitkorrelation (Rangkorrelation nach Spearman) und dem Reaktionsabstand der Metabolite im theoretischen Netzwerk (VGL1-Modellierung unter KEGG-Bedingungen) auf Basis der Glucose-Fermentation. Ausschließliche Betrachtung der exponentiellen Wachstumsphase. Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung. Anzahl der Datenpunkte: 1023. Interpolierte Dichtefunktion.

Konsequenzen: Eine zu Beginn dieser Arbeit gestellte Fragestellung, ob es einen übergeordneten, für alle Paarungen gültigen Zusammenhang zwischen der Prozessähnlichkeit zweier Konzentrationszeitreihen und deren korrespondierendem Reaktionsabstand im metabolischen Netzwerk gibt, kann auf der Basis der Ergebnisse eindeutig verneint werden. Ein Zusammenhang konnte auch nicht unter Verwendung anderer experimenteller Deskriptoren (wie zum Beispiel Winkelmaß, Gleichläufigkeit) festgestellt werden. Um diesen Sachverhalt tiefer zu beleuchten und um die gefundenen Teilergebnisse gegebenenfalls weitergehend zu verifizieren, wurden als nächstes nur die signifikant korrelierten Paarungen untersucht.

Hierzu wurde für die Korrelation ein Signifikanzniveau ($p<0{,}0001$) angelegt. Am Beispiel des betrachteten Datensatzes (Glucose-Fermentation, VGL1-Modellierung unter KEGG-Bedingungen), erfüllen nur 77 Metabolitkombinationen die gesetzten Signifikanzkriterien, was ungefähr einem Anteil von 7,5% an der verfügbaren Grundgesamtheit entspricht. Die Grundüberlegung für die Beschränkung auf signifikante Korrelationen lag darin, dass unter Umständen Ähnlichkeiten zwischen Metaboliten berechnet werden, die zwar theoretisch konnektiert sind, zwischen denen aber in Wirklichkeit unter den gegebenen Fermentationsbedingungen kein Stoffffluss existiert.

Die nachfolgende Grafik 7.19 beschäftigt sich mit den signifikant korrelierten Paarungen; auch hier wurde analog der Vergleich zum Reaktionsabstand durchgeführt. Neben der Spearman'schen Korrelation ist zusätzlich das Winkelmaß als Deskriptor dargestellt. In diesem reduzierten Datensatz waren fast ausschließlich positive Korrelationen zu finden (entspricht einem Anteil von 94,8%) , allenfalls 4 Antikorrelationen konnten festgestellt werden. Die negativen und positiven Zusammenhänge werden im weiteren Vorgehen separat behandelt. Die positiven Korrelationen sind in Abbildung 7.19 links dargestellt, die Antikorrelationen rechts.

Sowohl für Korrelation als auch für Winkelähnlichkeit zeigt sich, dass zumindest in der Tendenz ein Zusammenhang zum Reaktionsabstand vorhanden ist. Dieser Zusammenhang fasert bei den Paarungen positiven Zusammenhanges (links) mit abnehmender Korrelation beziehungsweise zunehmendem Winkelabstand weiter aus, wobei eine geringere Streuung bei Verwendung des Winkelmaßes festzustellen ist. Ferner führt beim Winkelmaß eine lineare Regression fast exakt durch den Nullpunkt. Die ähnlichsten hierbei gefundenen Metabolitpaare finden sich zwar in direkter Nachbarschaft, jedoch beträgt der größte Reaktionsabstand, welcher im

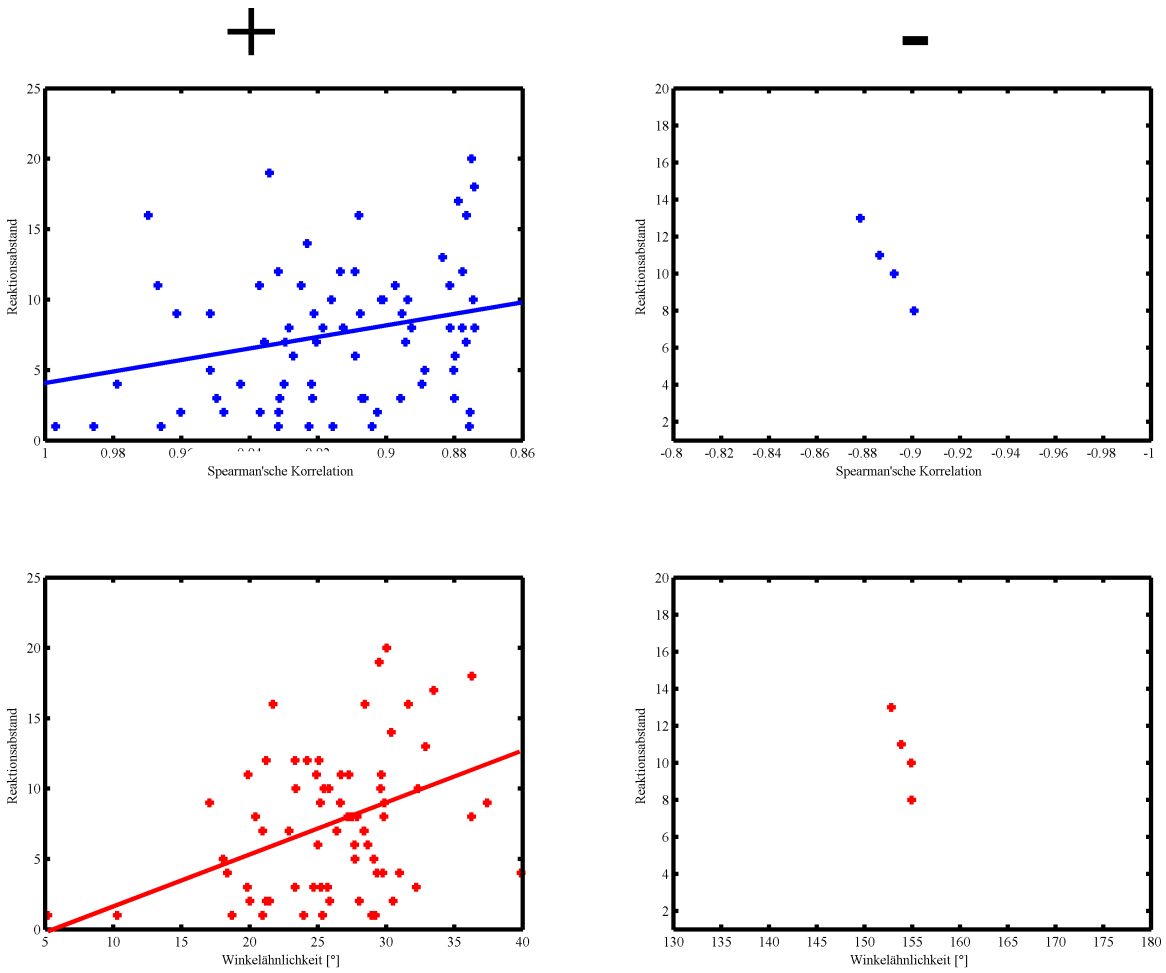

Abbildung 7.19: Betrachtung der Untergruppe signifikant ($p < 0{,}0001$) korrelierter Metabolitzeitreihen. Die Spearman'sche Korrelation als auch die Winkelähnlichkeit ist in Abhängigkeit des korrespondierendem Reaktionsabstandes dargestellt. Positive (n=73) und negative Korrelationen (n=4) wurden separat betrachtet. Datengrundlage: Glucose-Fermentation, sowie VGL1-Modellierung unter KEGG-Bedingungen. Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen. Experimentelle Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung.

betrachteten signifikanten Subset zu finden ist, immer noch 21 Reaktionsschritte. Die potenziellen Ursachen für signifikante Korrelationen unter großen Reaktionsabständen, bei denen keine direkte enzymatisch gesteuerte Umwandlung ursächlich sein kann, werden in Kapitel 7.4.2.6 diskutiert. Auch jenes Phänomen, dass benachbarte Metaboliten nicht immer die höchste Prozessähnlichkeit aufweisen müssen, findet sich im reduzierten Datensatz deutlich. Auch über die Ursachen dieses Phänomens soll an späterer Stelle diskutiert werden.

Betrachtet man die antikorrelierten Paarungen (rechts), so kann auch hier festgestellt werden, dass sie nur in relativ großen Reaktionsabständen zueinander zu finden sind. Obwohl die Stärke der Antikorrelation zwar tendenziell mit geringerem Reaktionsabstand zunimmt kann Aufgrund der geringen Anzahl von Datenpunkten an dieser Stelle keine belastbare Aussage für den Zusammenhang zum Reaktionsabstand getroffen werden. Offensichtlich scheint hingegen zu sein, dass signifikanten Antikorrelationen ein anderer Mechanismus zugrunde liegt als positiven Prozessähnlichkeiten. Über die möglichen Ursachen sei an späterer Stelle ausführlich diskutiert.

Konsequenzen: Bei Betrachtung der signifikant korrelierten Paarungen war es möglich einen (wenn auch schwachen) Zusammenhang zwischen der paarweisen Prozessähnlichkeit und korrespondierendem Reaktionsabstand der untersuchten Metaboliten festzustellen. Die sich am ähnlichsten Paare befinden sich tendenziell eher in kürzeren Abständen voneinander. Trotzdem finden sich nach wie vor signifikant korrelierte Paarungen in großen Reaktionsabständen. Signifikante Antikorrelationen konnten nur bei hohen Reaktionsabständen, nie in direkter Nachbarschaft im metabolischen Netzwerk beobachtet werden. Beachtet werden muss allerdings, dass in diesem Fall die Aussagen nur für einen Teil des gesamten Datensatzes gelten, von daher nicht verallgemeinert werden können. Zusätzliche Deskriptoren, wie zum Beispiel das Gibbs-Potenzial wurden zusätzlich für die Untersuchung des bestehenden Datensubsets herangezogen (siehe nächstes Kapitel). Ziel dieses Vorgehens war es, zu überprüfen, ob die Hinzunahme zusätzlicher Deskriptoren die festgestellten Zusammenhänge tiefergehend erläutert.

7.4.2.2 Prozessähnlichkeit, theoretische Pfadlänge und Gibbs-Potenzial

Um den festgestellten Zusammenhang zwischen Winkelmaß und Reaktionsabstand auf dem Subset signifikanter Ähnlichkeiten weiter zu hinterfragen, wurde mit dem Gibbs-Potenzial eine zusätzliche Variable ins Spiel gebracht. Das Gibbs-Potenzial wurde wie in Kapitel 5.2.4.5 beschrieben, für alle Reaktionen des theoretischen Netzwerkes abgeleitet. Abbildung 7.20 zeigt, wie sich das Maximum der Gibbs-Energie innerhalb eines modellierten metabolischen Pfades zur Winkelähnlichkeit und dem Reaktionsabstand auf Basis des erzeugten Datensubsets verhält.

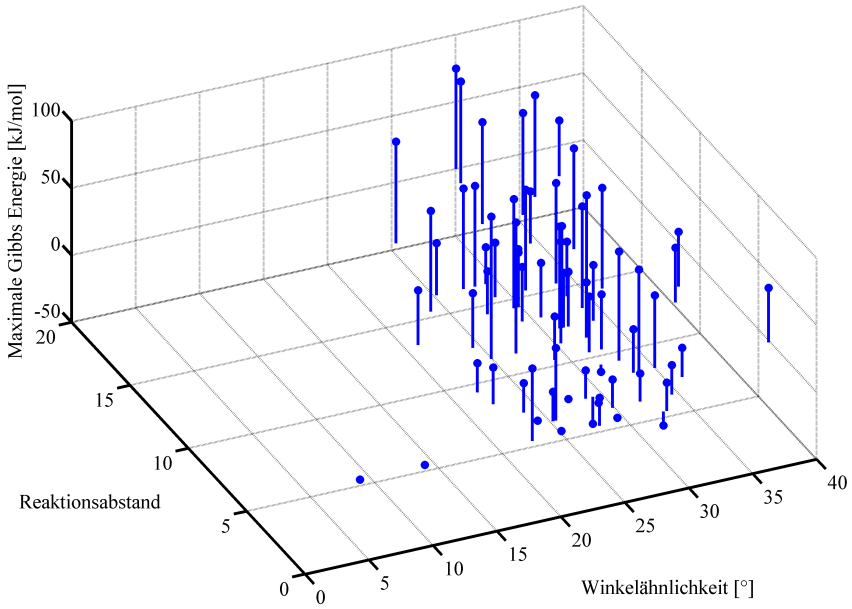

Abbildung 7.20: Zusammenhang zwischen theoretischer Pfadlänge, Winkelähnlichkeit und der maximalen Gibbs-Energie entlang des Pfades auf Basis der signifikanten korrelierten Paarungen ($p < 0{,}0001$). Daten aus der Glucose-Fermentation, sowie VGL1-Modellierung unter KEGG-Bedingungen. Theoretische Parameter: Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen. Experimentelle Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung.

Interessanterweise weisen jene Paare, die einen geringen Reaktionsabstand und

einen geringen Winkelabstand zueinander aufweisen, ein niedriges beziehungsweise negatives Gibbs-Potenzial auf. Das bedeutet: die betreffenden Reaktionen laufen entweder spontan unter Energiegewinnung ab, oder können gut durch enzymatische Aktivität reguliert werden. Mit zunehmendem Reaktionsabstand und zunehmender Winkelähnlichkeit finden sich im Allgemeinen größere positive Gibbs-Potenziale. Als Deskriptor wurde bewusst das Maximum des Gibbs-Potenzials entlang des metabolischen Pfades gewählt, da eine Mittelwertsbildung oder die Wahl des Minimums verschleiern würde, ob sich gegebenenfalls entlang des Pfades ein energetisch kostenintensiver Schritt befindet.

Konsequenzen: Die Hinzunahme der Gibbs-Energie ergab, dass für die ähnlichsten Paarungen in direkter Nachbarschaft negative Potenziale oder Potenziale nahe Null vorhanden waren. Obwohl die reaktionsspezifische Berechnung der Gibbs-Energie (vergleiche Kapitel 5.2.4.5) auf der Basis zahlreicher Annahmen beruht, könnte dies ein Hinweis darauf sein, dass die Reaktion zwischen den betreffenden Metaboliten entweder spontan, beziehungsweise durch enzymatische Aktivität katalysiert, abläuft. Für das Vorkommen signifikant hoher Prozessähnlichkeiten in großen Netzwerkabständen liefert auch die Betrachtung der Gibbs-Energie keinen Erklärungsansatz.

Aus dieser Erkenntnis heraus wurde abgeleitet, dass das Auftreten hoher Prozessähnlichkeiten in metabolischen Netzwerk auch losgelöst vom dazwischenliegenden Reaktionsabstand betrachtet werden muss. Zwei Ansätze wurden getestet: erstens sollte überprüft werden, ob hohe Prozessähnlichkeiten vom Konzentrationsverhältnis beider Metaboliten abhängig sein kann. Zweitens wurde Augenmerk auf die Tatsache gelegt, dass die Metaboliten innerhalb ihrer Netzwerke unterschiedliche Topologien besitzen, beispielsweise beschrieben durch die Anzahl der Nachbarn, mit denen sie durch Reaktionen in Wechselwirkung stehen (vergleiche hierzu Kapitel 7.4.1.1). Daraus resultierend wurde intensiv untersucht, ob eine hohe Prozessähnlichkeit von den paarweisen Nachbarschaftsbeziehungen der betrachteten Metaboliten abhängig ist.

7.4.2.3 Prozessähnlichkeit und paarweise Konzentrationsverhältnisse

In dieser Untersuchung wurde das Konzentrationsverhältnis zweier Metaboliten zum Zeitpunkt des höchsten Wachstums in der exponentiellen Wachstumsphase

mit seiner Prozessähnlichkeit verglichen. Da die Berechnung der Prozessähnlichkeit abhängig von der Vorverarbeitung der Daten ist, wurde ein Ähnlichkeitsindex entwickelt. Dieser Indikator besagt, wie oft unter Betrachtung aller Varianten der Datenvorverarbeitung eine definierte Metabolitpaarung eine signifikante Spearman-Korrelation mit einem P-Wert $<$ 0,0001 erreichte. Die Mehrzahl der Paarungen weist - wie bereits erwähnt - egal unter welcher Datenvorverarbeitung keinen derart starken statistischen Zusammenhang auf. Maximal ist ein Wert von 100 möglich, was bedeutet, dass die signifikant korrelierte Metabolitpaarung in allen Varianten der Datenvorverarbeitung zu finden ist.

Für den Zusammenhang von Konzentrationsverhältnis und Ähnlichkeitsindex konnte, wie in Abbildung 7.21 dargestellt, gezeigt werden, dass Metabolitpaare sehr hoher Prozessähnlichkeit, sich hinsichtlich ihrer Konzentrationsverhältnisse maximal um den Faktor 10 unterscheiden. Bei Paarungen von Metaboliten hingegen, die keine signifikanten Korrelationen aufweisen (also einen P-Wert größer 0,0001 besitzen, was einem Ähnlichkeitsindex von 0 entspricht) konnten Konzentrationsunterschiede von mehr als 5 Größenordnungen festgestellt werden. Dies kann darauf hindeuten, dass signifikante Korrelationen dann nicht auftreten können, wenn beide Metaboliten in einem stark unterschiedlichen Konzentrationsverhältnis vorkommen. Dass die Paarungen höchster Prozessähnlichkeit in vergleichbar ähnlicheren Konzentrationsverhältnissen vorkommen, könnte dafür sprechen, dass die betrachteten Metaboliten mehr oder weniger linear und von anderen Prozessen unbeeinflusst ineinander umgewandelt werden. Der Zusammenhang des Konzentrationsverhältnisses zur Prozessähnlichkeit konnte nicht bei allen Fermentationen gleichermaßen beobachtet werden. Sie war besonders deutlich bei der Fructose-Fermentation, und mit einigen Einschränkungen auch bei der Glucose- und Glutamin-Fermentation feststellbar. Interessanterweise war kein Zusammenhang bei der Fütterung mit Acetat und Lactat, welche vermutlich eher ähnlich verstoffwechselt werden, zu beobachten.

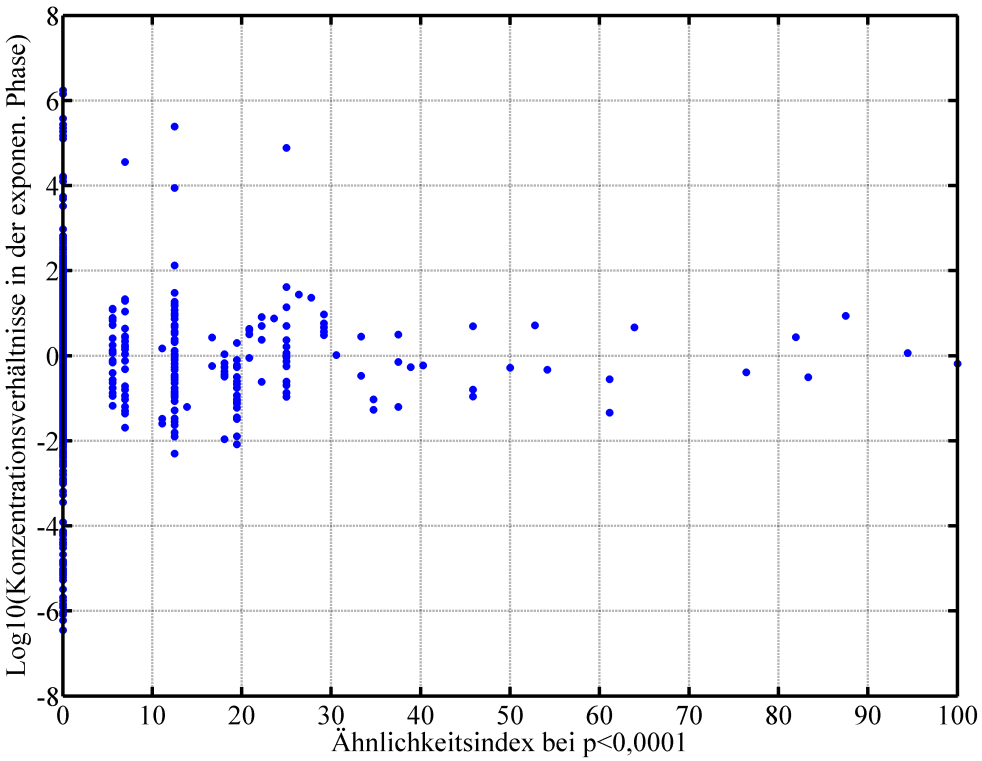

Abbildung 7.21: Zusammenhang zwischen dem Ähnlichkeitsscore der Metabolitpaarungen und ihren Konzentrationsverhältnissen, basierend auf den logaritmierten Konzentrationen der exponentiellen Wachstumsphase. Betrachtete Fermentation: Fructose, Reaktionsnetzwerk: VGL2. Theoretische Parameter: Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit, sowie Verwendung des KEGG-Mapping. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen. Experimentelle Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung.

7.4.2.4 Prozessähnlichkeit und paarweiser Verknüpfungsgrad

Stellt man den Ähnlichkeitsindex (welcher gewissermaßen als Maß für die Robustheit der paarweisen Prozessähnlichkeit in Bezug auf die verschiedenen Varianten der Datenvorverarbeitung dient) der mittleren Verknüpfungsgrad der Paarung gegenüber, so ergibt sich ein äußerst interessanter Sachverhalt. Der mittlere Verknüpfungsgrad der Paarung setzt sich hierbei aus dem Mittelwert der Verknüpfungsgrade beider Metaboliten (vergleiche Metabolitverknüpfung, Kapitel 5.2.4.6) zusammen.

Die Abbildung 7.22 zeigt, dass jene Metabolitpaare, welche eine ausgesprochen hohe Prozessähnlichkeit zueinander besitzen, jeweils vergleichsweise wenige Verknüpfungen zu anderen Metaboliten aufweisen. Bei jenen Metabolitpaarungen, welche einen niedrigen Ähnlichkeitsindex aufweisen und die gesetzten Signifikanzschwellen nicht erreichen, finden sich deutlich höhere Verknüpfungszahlen (bis über 20). In der Zusammenschau betrachtet bedeutet dies, dass höhere Prozessähnlichkeiten dann wahrscheinlich sind, wenn die beteiligten Metaboliten ihrerseits geringe Verknüpfungsgrade zu anderen Metaboliten aufweisen. Dieser Zusammenhang ist unabhängig von dem tatsächlichen Reaktionsabstand zwischen den betrachteten Metaboliten, sondern nur von deren Nachbarschaftsverhältnissen abhängig. Interessanterweise findet sich dieser Zusammenhang fermentationsübergreifend auch bei allen anderen Ausgangssubstraten (Acetat, Lactat, Fructose, sowie eingeschränkt bei Glutamin). Dies könnte als Hinweis darauf gewertet werden, dass es sich bei dem gefundenen Zusammenhang um ein generelles Phänomen handelt.

Konsequenzen: Die Betrachtung der paarweisen Prozessähnlichkeiten mit den Verknüpfungsgraden der korrespondierenden Metaboliten ergab, dass die Topologie des Netzwerkes entscheidenden Einfluss auf die beobachtbaren signifikanten Prozessähnlichkeiten hat. Sie können vorzugsweise dann beobachtet werden, wenn die betrachteten Metaboliten ihrerseits geringe Verknüpfungsgrade aufweisen. Eine hohe Korrelation zwischen einem Metabolit-Hub und einem anderen Metaboliten kann folglich als eher unwahrscheinlich angesehen werden. Für hohe Prozessähnlichkeit zwischen Metaboliten bei großen Reaktionsabständen könnte dies ein Hinweis darauf sein, dass beide Metaboliten nicht durch enzymatische Aktivität in Verbindung stehen, sondern im Gegenteil, dass beide Metaboliten -

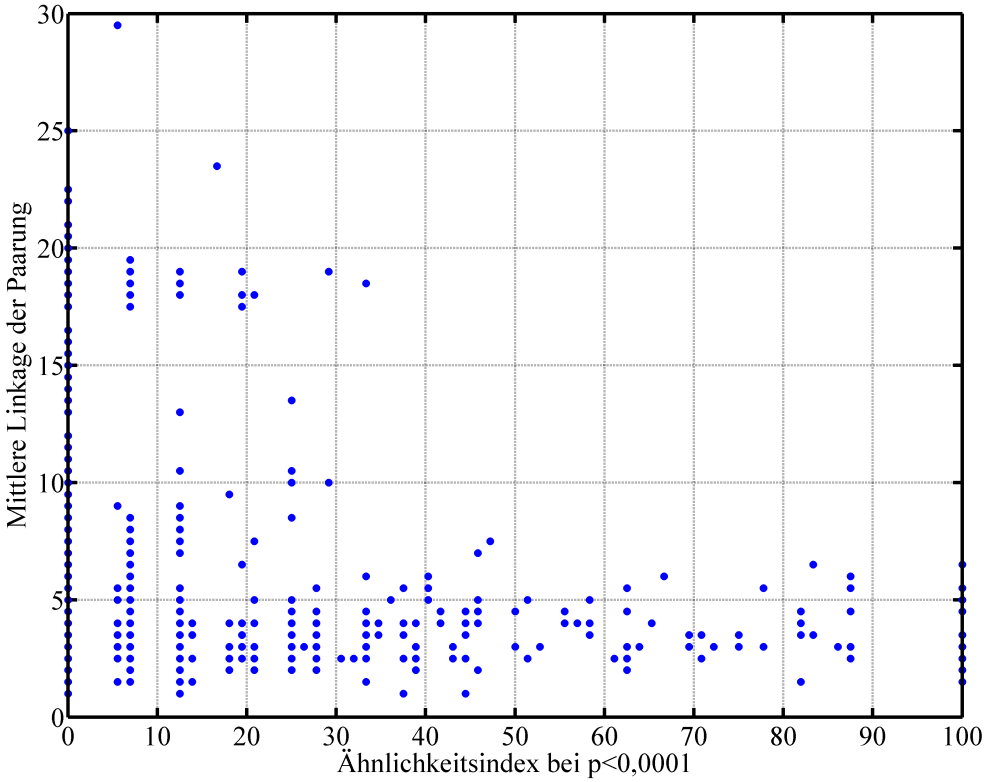

Abbildung 7.22: Zusammenhang zwischen dem Ähnlichkeitsscore der Metabolitpaarungen und dem mittleren paarweisen Verknüpfungsgrad. Daten aus der Glucose-Fermentation, sowie VGL1-Modellierung unter KEGG-Bedingungen.Theoretische Parameter: Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen. Experimentelle Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung.

relativ isoliert von weiteren Prozessen - einer übergeordneten Steuerung folgen. Mögliche Beispiele für solche Phänomene werden im zusammenfassenden Kapitel 7.4.2.6 gegeben.

7.4.2.5 Mittlere Prozessähnlichkeit und Verknüpfungsgrad

Kehrt man die Betrachtungsweise um und vergleicht nun die mittlere Prozessähnlichkeit eines Metaboliten zu seinen theoretischen Nachbarn, so sollte der im vorangegangenen Kapitel festgestellte Effekt deutlich sichtbar sein. Um dies zu überprüfen, wurde das Mittel aller Korrelationen eines Metaboliten zu seinen theoretisch verknüpften Nachbarn berechnet. Um die Korrelationswerte vergleichbar zu machen, wurden sie zur Mittelwertsberechnung z-transformiert und der gemittelte Wert anschließend retransformiert.

Die Abbildung 7.23 zeigt, dass die mittlere Pearson'sche Korrelation stark verknüpfter Metaboliten nahe Null liegt. Dies könnte die Ursachen darin haben, dass Metabolit-Hubs mit vielen Nachbarn durch enzymatische Aktivität in Verknüpfung stehen. Diese Verknüpfungen können theoretisch gesehen sowohl einen positiven als auch einen negativen Zusammenhang aufweisen, sodass die Vermutung nahe liegt, dass sie sich im Mittel ausgleichen und deshalb eine mittlere Prozessähnlichkeit im unkorrelierten Bereich erzeugen. Die Annahme, dass sich die Einflüsse der zahlreichen Nachbarn im Mittel aufheben, konnte auch hinsichtlich der Untersuchung der Konzentrationen in Zusammenschau mit dem Verknüpfungsgrad (vergleiche Kapitel 7.4.1.1) angeführt werden. Die Konzentration von Metabolit-Hubs, wie beispielsweise Pyruvate (C00022), blieben in den betrachteten Fermentationsexperimenten häufig sehr konstant.

Bei den schwach verknüpften Metaboliten zeigt sich, dass die mittlere Prozessähnlichkeit in der Mehrzahl deutlich höhere Wertebereiche annimmt. Dies kann darauf hindeuten, dass die Metaboliten stärker von ihren wenigen Nachbarn abhängig sind. Die gemittelten Korrelationen sind im Allgemeinen positiv, nur sehr wenige antikorrelierte Paare sind zu beobachten. Stellt man sich schwach verknüpfte Metaboliten in Form einer gerichteten Kette angeordnet vor, so könnte eine hohe Korrelation (besonders unter Berücksichtigung eines Reaktionsradius von 2 Schritten) darauf hindeuten, dass der Metabolit in seinem Prozessverhalten sowohl seinem Vorläufer als auch seinem Nachfolger ähnelt und dass weitere Einflüsse, wie zum Beispiel Abzweigungen, nicht existieren.

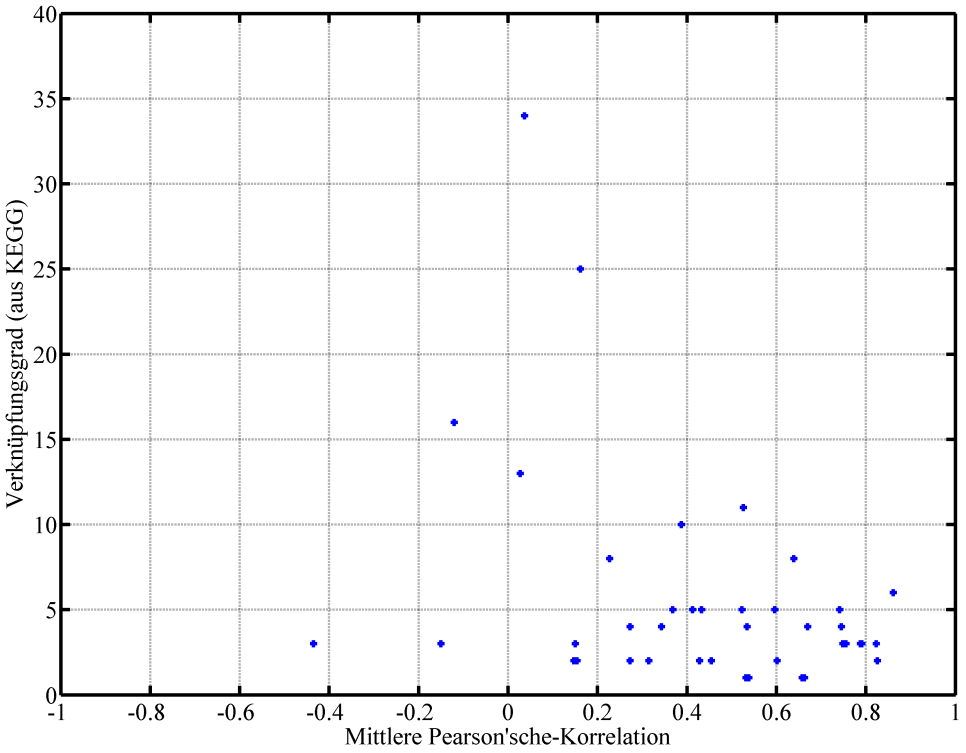

Abbildung 7.23: Zusammenhang zwischen der mittleren Prozessähnlichkeit und dem theoretischen Verknüpfungsgrad im Radius von zwei Reaktionsschritten. Daten aus der Glucose-Fermentation, sowie VGL1-Modellierung unter KEGG-Bedingungen. Theoretische Parameter: Schwellenwerte von 15% lokaler und 1% globaler Molekül-Ähnlichkeit. Reaktionswege über Seitenmetaboliten, wie in Kapitel 4.3.3 beschrieben, wurden nicht zugelassen. Experimentelle Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung.

7.4.2.6 Zusammenfassende Betrachtung zu temporalen Prozessähnlichkeiten in metabolischen Netzwerken

Die wichtigsten Erkenntnisse bei der Betrachtung von paarweisen Prozessähnlichkeiten zwischen Metaboliten werden an dieser Stelle zusammengefasst und Erklärungsmöglichkeiten für die beobachteten Phänomene gegeben.

- Signifikant korrelierte Konzentrationszeitreihen von Metaboliten existieren in direkter theoretischer Nachbarschaft als auch großem Reaktionsabstand voneinander. Befinden sie sich in direkter Nachbarschaft, so ist es wahrscheinlich, dass sie in einem - enzymatisch katalysiertem - Zusammenhang stehen. Eine Erklärungsmöglichkeit für hohe Prozessähnlichkeiten in großen Reaktionsabständen konnten Paarungen zwischen Aminosäuren liefern. Beispielsweise konnten hohe Ähnlichkeiten zwischen L-Homoserine (C00263) und L-Valine (C00183) gefunden werden, welche im metabolischen Netzwerk relativ weit (9 Reaktionschritte in der VGL1-Modellierung unter KEGG-Bedingungen) voneinander entfernt sind. Dies ist nur ein exemplarisches Beispiel. Hohe Prozessähnlichkeit findet sich relativ häufig zwischen Aminosäuren, es konnte auch ferner fermentationsübergreifend festgestellt werden, dass sie häufig dem gleichen Cluster zugeordnet werden (vergleiche Kapitel 7.1.4.2). Eine parallele, bedarfsgerechte und hierarchisch gesteuerte Produktion der Aminosäuren könnte eine Ursache der hohen Korrelation sein.

- Metabolitzeitreihen hoher Prozessähnlichkeit konnten nicht beobachtet werden, wenn sich die Konzentrationsverhältnisse der Metaboliten stark unterscheiden. Die ähnlichsten Paarungen finden sich dann, wenn sich die Konzentrationsverhältnisse (in der exponentiellen Phase) nicht mehr als eine Größenordnung voneinander unterschieden.

- Unabhängig vom Reaktionsabstand treten signifikant hohe Prozessähnlichkeiten vorzugsweise dann auf, wenn die betrachtete Metabolitpaarung ihrerseits einen vergleichsweise geringen mittleren Verknüpfungsgrad aufweist. Diese Erkenntnis steht in Einklang mit der möglichen parallelen Produktion ausgewählter Aminosäuren.

- Stark verknüpfte Metaboliten weisen eine mittlere Korrelation um Null auf, wenn die Prozessähnlichkeit zu den theoretischen Nachbarn betrachtet wird.

Weniger stark verknüpfte Metaboliten hingegen weisen deutlich höhere gemittelte Korrelationen auf, was dafür spricht, dass die Position sowie die Verknüpfung des Metaboliten im metabolischen Netzwerk entscheidend die paarweise Prozessähnlichkeit beeinflusst.

- Antikorrelierte Metaboliten konnten in den beobachteten Daten nicht in direkter Nachbarschaft gefunden werden. Auch für dieses Phänomen kann im Rahmen dieser Arbeit eine Erklärungsmöglichkeit gegeben werden. Abzweigungen im metabolischen Netzwerk können theoretisch dergestalt reguliert sein, dass die Präsenz eines Metaboliten direkt die Synthese eines anderen Metaboliten - beispielsweise durch Inhibition des katalysierenden Enzyms - beinflusst. Exemplarisch kann der Zucker / Stärke-Stoffwechsel angeführt werden, wie er bei Bäumen unter winterlichen Bedingungen existiert. Je höher hier der Saccharose-Gehalt ist, desto geringer ist der Stärkegehalt und umgekehrt. Beide Metaboliten befinden sich auch nicht in direkter Nachbarschaft, sondern sind einige Reaktionsschritte voneinander entfernt. Ihrer Synthese liegen folglich gegenläufige Prozesse zugrunde. Solch gegenläufige Prozesse können unter Umständen auch die Ursache der antikorrelierten Zeitreihen dieser Arbeit sein.

- Die betrachteten theoretischen metabolischen Netzwerke stellen statische Informationen dar. Sie geben zwar an, ob und wie beispielsweise zwei Metaboliten, enzymatisch katalysiert, ineinander überführt werden können, aber sie sagen nichts darüber aus, ob das entsprechende Enzym überhaupt unter den gegebenen Bedingungen in der Zelle vorhanden ist. Informationen darüber erhält man, indem man die Expression der enzymkodierenden Gene untersucht. Dies wurde im nachfolgenden Unterkapitel analysiert.

7.4.3 Substratspezifische Untersuchung von Metabolomdaten, theoretischen Netzwerktogologien und Transkriptominformationen

Bei allen betrachteten Ausgangssubstraten gelingt es *Corynebacterium glutamicum*, sämtliche essentiellen Bausteine zu synthetisieren und somit sein Überleben zu sichern. Dass dies in Abhängigkeit vom vorhandenen Substrat nicht immer gleichartig und gleichermaßen effizient funktioniert, zeigte sich in vorangegangenen Untersuchungen. Um festzustellen, welche Gene eine maßgebliche Rolle bei dieser Anpassungsfähigkeit in Bezug auf verfügbare Ausgangssubstrate (Eggeling und Bott, 2005) spielen, wurden in den letzten Jahren an *C. glutamicum* eine große Anzahl von Untersuchungen durchgeführt, welche die Genexpression mit Hilfe der Microarray-Technologie bestimmten. Einige dieser Untersuchungen wurden unter vergleichbaren Fütterungsbedingungen wie in dieser Arbeit durchgeführt und konnten daher herangezogen werden. In dem nun folgenden Kapitel wird untersucht, ob sich substratinduzierte Unterschiede in der Expression enzymkodierender Gene auch in der Prozessähnlichkeit benachbarter Metabolitzeitreihen widerspiegeln.

7.4.3.1 Differenzielle Untersuchung von Transkriptom und Metabolom unter Fütterungsbedingungen mit Glucose und Acetat

Die Unterschiede in der Expression enzymkodierender Gene unter Fütterungsbedingungen von *C. glutamicum* mit Glucose sowie Acetat wurden intensiv von mehreren Forschergruppen untersucht. Die wichtigsten Unterschiede in der transkriptionellen Aktivität von *C. glutamicum* in Abhängigkeit von den verwendeten Ausgangssubstraten Glucose und Acetat werden nachfolgend am Beispiel des Zentralstoffwechsels detailliert diskutiert.

Wenn Acetat als Nährstoffquelle zur Verfügung steht, zeigt sich, dass die Expression enzymkodierender Gene der Glykolyse und des Pentose-Phosphat-Weges signifikant im Vergleich zu Fütterungsbedingungen mit Glucose heruntergeregelt ist (Hayashi et al., 2002; Muffler et al., 2002 und Gerstmeir et al., 2003). Der Grund dafür liegt darin, dass Acetat nicht, wie bei der Fütterung mit Glucose über die oben genannten Stoffwechselwege, sondern primär über Acetyl-CoA (C00024) aufgenommen und in den Zitratzyklus eingespeist wird; eine auffällige differenzielle

Exprimierung der hierfür verantwortlichen enzymkodierenden Gene *ack* und *pta* konnte in den Transkriptomuntersuchungen ebenfalls nachgewiesen werden. Auch die Untersuchungen von Wendisch et al. (2000), welche *C. glutamicum* auf Glucose und Acetat untersuchten und mit Hilfe radioaktiv markierter ^{13}C-Isotope die metabolischen Stoffflüsse quantifizierten, zeigten, dass Acetat über den oben genannten Weg aufgenommen wird (vergleiche hierzu auch Abbildung 7.24, welche die oben genannten Ergebnisse zusammenfasst). Auch im Zitratzyklus, in dem sowohl katabolische als auch anabolische Prozesse gleichermaßen ablaufen können und dessen Aktivität für das Überleben von Organismen unabdingbar ist (Krebs und Johnson, 1937), zeigen sich bei *C. glutamicum* deutliche substratinduzierte Unterschiede in der Expression enzymkodierender Gene. So lässt sich beispielsweise unter Acetat-Bedingungen eine Besonderheit im Zitratzyklus feststellen, die als Glyoxylat-Kurzschluss („Glyoxylate-Shunt“) bezeichnet wird. Hierbei handelt sich um eine Abkürzung innerhalb des Zitratzyklus, welche vom Metaboliten Isocitrate zu Succinate respektive Malate reicht (vergleiche Abbildung 7.24). Die für diesen Kurzschluss verantwortlichen Gene *aceA* und *aceB* sind ebenfalls signifikant im Vergleich zu Glucose-Bedingungen exprimiert. Auch für alle anderen enzymkodierenden Gene des Zitratzyklus gilt, dass sie unter Acetat-Bedingungen eine stärkere Exprimierung aufweisen (Gerstmeir et al., 2003). Dies betrifft im Einzelnen die Gene: *acn*, *gltA*, *sdhA*, *sdhB*, *sdhCD*, *fumH*, und *mdh*. Die höhere Aktivität des Zitratzyklus unter Fütterungsbedingungen mit Acetat ist dadurch zu erklären, dass dieser primär zur Energiegewinnung und zur Produktion von Vorläufermetaboliten genutzt werden muss. Vergleicht man nun substratabhängig die Prozessähnlichkeit der Zeitreihen direkt benachbarter Metaboliten in Zusammenschau mit der Expression jener Gene, welche die entsprechenden katalysierenden Enzyme kodieren, so lassen sich einige äußerst interessante Auffälligkeiten feststellen. Diese Ergebnisse werden in den folgenden Unterkapitel zusammengefasst und in Abbildung 7.24 grafisch aufgearbeitet. Ein wichtiger Hinweis vorab bezüglich der erwähnten Abbildung. In ihr sind aus Gründen der Übersichtlichkeit die Pentose-Phosphate gruppiert dargestellt, ferner wurden Abkürzungen für die Metabolitbezeichnungen benutzt. Folgende Abkürzungen sind in der Grafik verwendet: Xu5P für D-Xylulose 5-phosphate, Ru5P für D-Ribulose 5-phosphate, R5P für D-Ribose 5-phosphate, GA3P für D-Glyceraldehyde 3-phosphate, 3PG für 3-Phospho-D-glycerate, 2PG für 2-Phospho-D-glycerate, PEP für Phosphoenolpyruvate und PYR für Pyruvate.

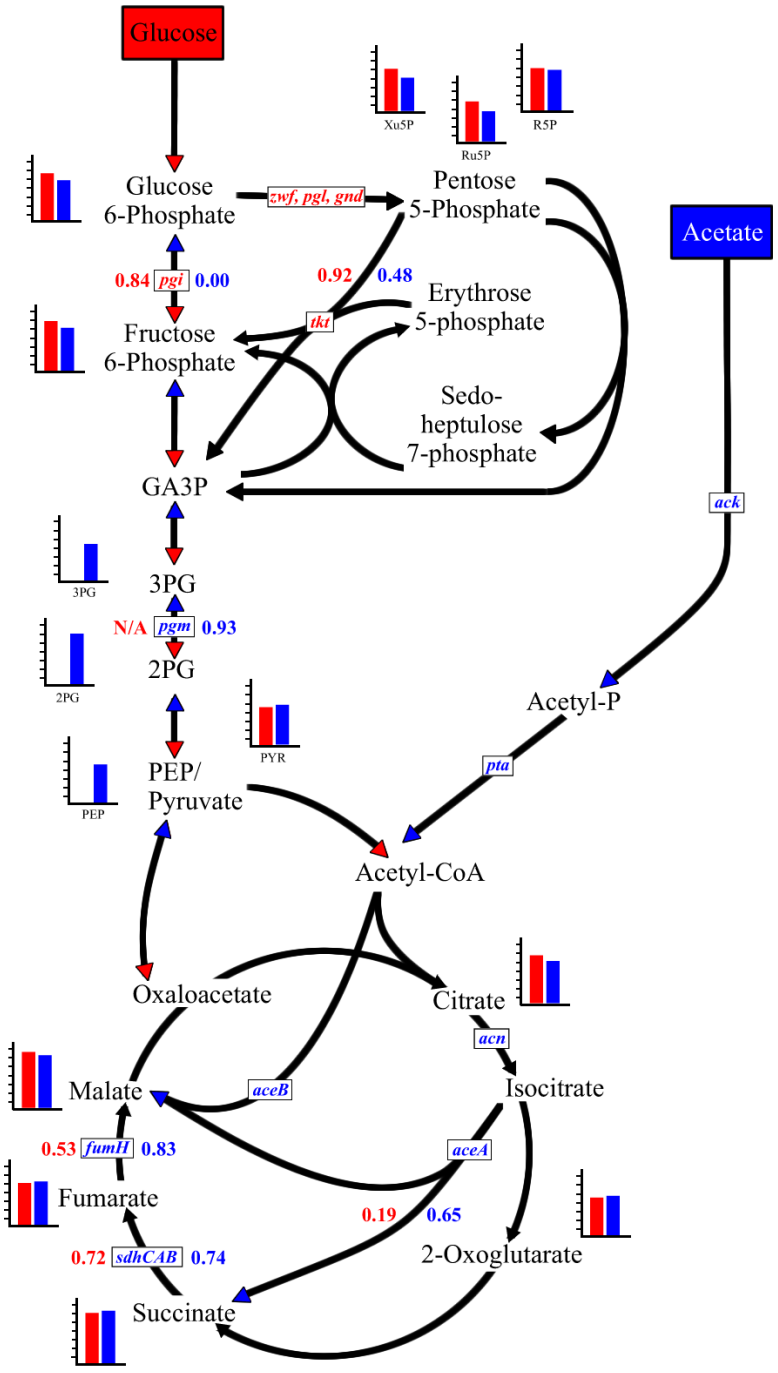

Abbildung 7.24: Schematische Darstellung des Zentralstoffwechsels von *C. glutamicum* unter Fütterungsbedingungen mit Glucose (rot) und Acetat (blau). Darstellung der mutmaßlichen Flussrichtungen, der Metabolitkonzentrationen in der exponentiellen Wachstumsphase (schematisierte Balkendiagramme), der exprimierten Enzymgene (kursiv) sowie der zugehörigen Prozessähnlichkeiten (abgegriffen durch Spearman'sche Korrelation). (Grafik angepasst nach den Arbeiten von Wendisch et al., 2000).

Im Einzelnen lassen sich folgende Auffälligkeiten feststellen:

- Die Umwandlung der Metaboliten Glucose 6-Phosphate (C00092) und Fructose 6-Phosphate (C00085) - einer der ersten Reaktionsschritte der Glykolyse - wird durch das Enzym Glucose-6-phosphat Isomerase (EC 5.3.1.9) katalysiert (Reaktionsschritt R00771). Beide Metaboliten zeigen unter Fütterungsbedingungen mit Glucose eine hohe Prozessähnlichkeit, welche sich in einem Korrelationskoeffizienten von r=0,84 äußert. Gleiche Metaboliten sind jedoch unter Fütterungsbedingungen mit Acetat in ihrem Prozessverhalten gänzlich unkorreliert (r=0,0). Das korrespondierende Enzym-Gen *pgi* ist bei Fütterungsbedingungen mit Glucose laut den Arbeiten von Muffler et al. (2002) und Hayashi et al. (2002) stärker exprimiert, als unter Fütterungsbedingungen mit Acetat, wenngleich diese Veränderungen nicht als signifikant erachtet worden sind. Die Arbeiten von Dominguez et al. (1998) postulieren hingegen, dass unter Fütterung mit Glucose die besagte Reaktion nahe am thermodynamischen Gleichgewicht operiert und deshalb unter diesen Umständen nicht enzymatisch kontrolliert werden kann. Es zeigt sich, dass diese Reaktion besonders detailliert in Zusammenschau mit der Topologie der betreffenden Metaboliten untersucht werden muss. Der Grund hierfür besteht insbesondere aus der Verbindung von Glucose 6-Phosphat zum Pentose-Phosphat-Weg, welche als wichtige Abzweigung im metabolischen Netzwerk anzusehen ist. Aus der Untersuchung des Transkriptoms ist ferner bekannt, dass die Expression von Enzymgenen, welche Glucose 6-phosphate in Richtung des Pentose-Phosphat-Weges weiter verstoffwechseln, ebenfalls signifikant erhöht ist. Aufgrund der Tatsache, dass manche Metaboliten entlang dieses Pfades messtechnisch nicht erfasst werden konnten, (wobei insbesondere der Metabolit D-Glucono-1,5-lactone 6-phosphate (C01236) zu nennen ist, welcher durch das Enzym Glucose-6-phosphate Dehydrogenase (EC 1.1.1.49) konnektiert ist), konnten die paarweisen Prozessähnlichkeiten entlang dieses Pfades nicht weitergehend bestimmt werden.

  Wie im nächsten Unterkapitel beschrieben wird, weisen die Transkriptomuntersuchungen ferner auf eine erhöhte Aktivität des Enzyms Transketolase (EC 2.2.1.1) hin. Betrachtet man all diese Informationen in Zusammenschau, so erscheint es denkbar, dass die hohe Prozessähnlichkeit zwischen Glucose 6-phosphate und Fructose 6-phosphate zum einen aus der direkten

Umsetzung als auch aus dem längeren, jedoch sehr aktiven Reaktionsweg in den Pentose-Phosphat-Weg und zurück in die Glykolyse bewirkt werden kann. Diese Annahme wird durch die Arbeiten von Wendisch et al. (2000) gestützt. In dieser Arbeit wurden die metabolischen Flussraten im Zentralstoffwechsel von *C. glutamicum* anhand radioaktiv markierter Isotope bestimmt. Ihre Ergebnisse zeigen, dass sowohl die direkte Umsetzung als auch der längere Reaktionsweg über die Transketolase gleichermaßen unter Fütterungsbedingungen mit Glucose aktiv sind. In diesem Zusammenhang muss erwähnt werden, dass die in dieser Arbeit ermittelten Flussraten für den längeren Reaktionsweg deutlich höhere Werte annehmen, als für die direkte enzymatische Umwandlung. Für den Vergleich zu Fütterungsbedingungen mit Acetat zeigte sich, dass generell nur sehr geringe Flussraten für metabolische Pfade von der Glykolyse in den Pentose-Phosphat-Weg und zurück existieren. Dieses Beispiel zeigt, dass die Betrachtung der Topologie paarweiser Metaboliten auch unter zusätzlicher Betrachtung von Transkriptomdaten äußerst sinnvoll ist und weiterführende Erklärungsmöglichkeiten liefert.

- Das bereits erwähnte Enzym Transketolase (EC 2.2.1.1), welche durch das Gen *tkt* kodiert wird, ist ein äußerst wichtiges Enzym, welches den Pentose-Phosphat-Weg mit der Glykolyse verbindet. Ihre Besonderheit besteht unter anderem darin, dass sie gleich mehrere Metaboliten miteinander enzymatisch verbindet (siehe auch obiges Beispiel und Abbildung 7.24). Eine dieser Verbindungen katalysiert die Transketolase beispielsweise zwischen den Metaboliten D-Xylulose 5-phosphate (C00231) und beta-D-Fructose 6-phosphate (C05345). Unter Fütterungsbedingungen mit Glucose ist das entsprechende Gen *tkt*, verglichen mit Bedingungen unter Acetat-Fütterung signifikant exprimiert. Die Zeitreihen der beiden Metaboliten zeigen in der Acetat-Fermentation mit einem Korrelationskoeffizienten von r=0,47 keine nennenswerte Auffälligkeit, während sie bei der Fütterung mit Glucose r=0,92 einen sehr deutlichen Zusammenhang aufweisen. Diese Erkenntnis stützt die aus den Transkriptomdaten abgeleitete Vermutung, dass Glykolyse und Pentose-Phosphat-Weg unter Fütterungsbedingungen mit Glucose deutlich aktiver sein müssen als unter Acetat-Bedingungen. Die aus den Isotopuntersuchungen abgeleiteten Flussraten von Wendisch et al. (2000)

zeigen für die entsprechende Reaktion unter Fütterungsbedingungen mit Glucose deutlich erhöhte Werte.

- In der Glykolyse findet sich unter Acetat-Bedingungen eine hohe Korrelation von r=0,93 zwischen den Metaboliten 2-Phospho-D-glycerate (C00631) und 3-Phospho-D-glycerate (C00197). Beide Metaboliten können durch das Enzym Phosphoglycerat Mutase (EC 5.4.2.1) reversibel ineinander überführt werden (R01518). Das entsprechende Enzym-Gen ist laut den Untersuchungen von Hayashi et al. (2002) unter Fütterungsbedingungen mit Acetat stärker exprimiert. Interessant ist hierbei, dass die gleichen Metaboliten unter Glucose-Bedingungen überhaupt nicht detektiert werden konnten, obwohl die Glykolyse nachweislich aktiv sein muss (Wendisch et al., 2000). Die Ursachen hierfür liegen außerhalb des Betrachtungssystems dieser Arbeit. Da Acetat, wie oben beschrieben, nicht durch die Glykolyse aufgenommen und verstoffwechselt wird, ist es wahrscheinlich, dass zwischen 2-Phospho-D-glycerate und 3-Phospho-D-glycerate unter Acetat-Bedingungen nicht Glykolyse sondern Glukoneogenese stattfindet, der metabolische Stoffffluss also „stromaufwärts“ zu den Zuckern führt. Diese Vermutung kann auch durch die Isotopenuntersuchung bei Fütterung von *C. glutamicum* mit Acetat durch die bereits erwähnten Arbeiten von Wendisch et al. (2000) bestätigt werden. Für den Fall der (wahrscheinlichen) Glukoneogenese unter Acetat-Bedingungen zeigen die Flussraten zwischen beiden Metaboliten deutlich geringere Werte als im Fall der Glykolyse, wenn sie unter Fütterungsbedingungen mit Glucose abläuft. Zusammengefasst könnte dies ein Hinweis darauf sein, dass die Glukoneogenese wahrscheinlich weniger komplex reguliert ist, als die Glykolyse.

- Der Zitratzyklus nimmt - wie bereits erwähnt - eine Schlüsselstellung im Stoffwechsel ein. Unter Acetat-Bedingungen, so zeigen betrachtete Untersuchungen des Transkriptoms einstimmig, muss der „Glyoxylate-Kurzschluss“, welcher gewissermaßen eine Abkürzung im Zitratzyklus darstellt, aktiv sein. Die Gene *aceA* und *aceB* kodieren das Enzym Isocitrat Lyase (EC 4.1.3.1) beziehungsweise das Enzym Malate Synthase (EC 2.3.3.9). Auch aus den Metabolitkonzentrationen lassen sich Hinweise ableiten, dass der „Glyoxylate-Kurzschluss“, unter Fütterungsbedingungen mit Acetat aktiv ist. Dies äußert sich in einer im Vergleich zu Glucose-Bedingungen erhöhten Prozess-

ähnlichkeit zwischen den Metaboliten von Citrate (C00158) und Succinate (Succinate). Obwohl Isocitrate (C00311) experimentell nicht detektiert werden konnte, zeigt es sich, dass die Konzentrationszeitreihen von Citrate und Succinate eine Korrelation r=0,65 besitzen, während sie bei der Glucose-Fütterung mit r=0,16 annähernd unkorreliert sind.

- Auch für weitere Metaboliten des Zitratzyklus konnten Unterschiede in der paarweisen Prozessähnlichkeit festgestellt werden. So unterscheidet sich beispielsweise die Korrelation zwischen den Metaboliten Fumarate (C00122) und (S)-Malate (C00149), welche katalysiert durch das Enzym Fumarate Hydratase (EC 4.2.1.2) ineinander überführt werden können, mit einem Koeffizienten von r=0,84 bei der Acetat-Fermentation deutlich von einem Korrelationskoeffizienten von r=0,52 für die Fütterung mit Glucose. Das korrespondierende Enzym-Gen *fumH* weist bei Fütterung von Acetat eine deutlich stärkere Exprimierung auf. Für die ebenfalls im Zitratzyklus benachbarten Metaboliten Succinate (C00042) und Fumarate (C00122) ist interessanterweise nur geringfügig erhöhte Prozessähnlichkeit strotz signifikant erhöhter Genexpression unter Fütterungsbedingungen mit Acetat festzustellen. Die Prozessähnlichkeit weist unter Acetat-Bedingungen einen Korrelationskoeffizienten von r=0,74 auf, während er unter Glucose-Bedingungen mit r=0,72 fast gleiche Wertebereiche annimmt. Das katalysierende Enzym für diesen Reaktionsschritt (R00412) ist Succinate Dehydrogenase (EC 1.3.99.1), welches durch das Gen *sdh* kodiert wird. Diese Paarung weist eine Besonderheit auf. Sie gehört, wie das nächste Kapitel 7.5 zeigen wird, zu den Paarungen, die sich hinsichtlich ihrer paarweisen Prozessähnlichkeit am wenigsten über alle betrachteten Fermentationen hinweg ändern.

Betrachtet man neben den paarweisen Prozessähnlichkeiten auch die Konzentrationsunterschiede der gemessenen Metaboliten (schematisch für die detektierten Metaboliten in Abbildung 7.24 durch die schematischen maßstabsgerechten Balkendiagramme der logarithmischen Konzentration dargestellt), so fällt auf, dass unter Fütterungsbedingungen mit Glucose Metaboliten wie Glucose 6-Phosphate (C00092), Fructose 6-Phosphate (C00085), sowie die Pentosephosphate D-Xylulose 5-phosphate (C00231), D-Ribulose 5-Phosphate (C00199) sowie D-Ribose 5-Phosphate (C00117) in deutlich erhöhter Konzentration verglichen mit den Acetat-Bedingungen auftreten. Dies ist insofern plausibel, da unter Acetat-

Bedingungen oben genannte Metaboliten überhaupt erst durch Glukoneogenese erreicht werden können. Im Zitratzyklus hingegen sind für die Metaboliten (S)-Malate (C00149) und Fumarate (C00122) höhere Konzentrationen unter Acetat-Bedingungen festzustellen, was mit der generell erhöhten Aktivität des Zitratzyklus als auch der Besonderheit des Glyxolat-Kurzschlusses in Zusammenhang stehen könnte.

Diese Ergebnisse zeigen deutlich, dass sich die Prozessähnlichkeit benachbarter Metaboliten deutlich in Abhängigkeit der Expression des entsprechenden enzymkodierendenden Gens ändert. Dies bedeutet, dass sich Vorgänge, welche auf der Ebene des Transkriptoms stattfinden, durchaus bis in die Zeitreiheneigenschaften von Metabolitkonzentrationen durchpausen können. Dies konnte bisher in keiner vergleichbaren Arbeit nachgewiesen werden. Auch ohne die direkte analytische Erfassung des Proteoms lassen sich somit erste Rückschlüsse über das Vorhandensein von Enzymen treffen.

Aus den Ergebnissen kann außerdem abgeleitet werden, dass für den betrachteten Organismus durch Anschalten spezialisierter Enzymgene die Möglichkeit gegeben ist, flexibel und schnell auf Umweltveränderungen (wie beispielsweise die Verfügbarkeit von Nährstoffen) zu reagieren. Durch verstärkte Produktion von Enzymen kann folglich der metabolische Fluss an - für den Organismus wichtigen - Stellen im Netzwerk gesteuert werden, was die Energieproduktion, die Synthese sämtlicher essentieller Bausteine sowie den Aufbau von Biomasse - kurzum, das Überleben - ermöglicht.

7.5 Paarweise Prozessähnlichkeit zwischen Metabolitzeitreihen als diskriminatorische Größe

Wie die Ergebnisse des vorangegangenen Kapitels eindrucksvoll zeigen, können sich Vorgänge, die auf der Ebene des Transkriptoms stattfinden, bis in die Eigenschaften experimentell erfasster Metabolitkonzentrationen durchpausen. Basierend auf dieser Erkenntnis wurden die experimentellen Deskriptoren aller Fermentationsexperimente in einer gemeinsamen Analyse untersucht. Dies bedeutet, dass die paarweisen Ähnlichkeiten aller Metabolitpaarungen aus allen Fermentationsexperimenten gemeinsam analysiert wurden. Insgesamt existieren für alle Experimente knapp 2000 individuelle Metabolitpaarungen. Da jedoch, wie bereits im Vorfeld erwähnt, nicht alle Metaboliten in allen Fermentationen nachgewiesen werden konnten, fand eine Gruppe von 854 gleichsam in allen Experimenten vorhandenen Metabolitpaarungen in dieser Analyse Berücksichtigung.

Für diesen gemeinsamen Datensatz wurde die Prozessähnlichkeit anhand der folgenden Deskriptoren untersucht: Pearson'sche Korrelation, Spearman Korrelation und Winkelähnlichkeit, wobei die Analyse separat für die verschiedenen Deskriptoren durchgeführt wurde. Ziel dieser Untersuchung war es, herauszufinden, ob sich substratinduzierte Unterschiede im Stoffwechsel von *C. glutamicum* in den berechneten Deskriptoren wiederfinden lassen. Um dies zu klären, wurde in einem ersten Schritt eine Hauptkomponentenanalyse auf den gemeinsamen Metabolitpaarungen für die berechneten Deskriptoren durchgeführt.

Abbildung 7.25 zeigt, dass sich die einzelnen Experimente deutlich hinsichtlich des verwendeten Ausgangssubstrat - basierend auf der Pearson'schen Korrelation als Deskriptor - trennen lassen. Die Wiederholungsexperimente weisen einen sehr geringen Abstand zueinander auf, während der Abstand zwischen den Substratgruppen deutlich größer ist. Die Hauptkomponente 1 - dargestellt auf der Abszisse - trennt in diesem Fall die Lactat- und Fructose-Fermentationen von den Fermentationen auf Acetat, Glutamin und Glucose. Hauptkomponente 2 trennt die Acetat- und Lactat-Fermentationen von den Fermentationen auf Glutamin, Fructose und Glucose. Eine ähnlich gute Trennung ergab sich ebenfalls für die Spearman'sche Korrelation und die Winkelähnlichkeit. Dieses Ergebnis kann dergestalt interpretiert werden, dass die Deskriptoren zur Beschreibung der Prozess-

ähnlichkeit einerseits in der Lage sind, charakteristische Merkmale aus den Konzentrationszeitreihen abzugreifen und ferner das Potenzial zur Trennung verschiedener substratinduzierter Veränderungen des Stoffwechsels besitzen. Vereinfacht ausgedrückt kann also gesagt werden, dass durch die Berechnung der paarweisen Prozessähnlichkeit von Metabolitzeitreihen die substratinduzierten Variationen des Metabolismus gut charakterisiert werden können.

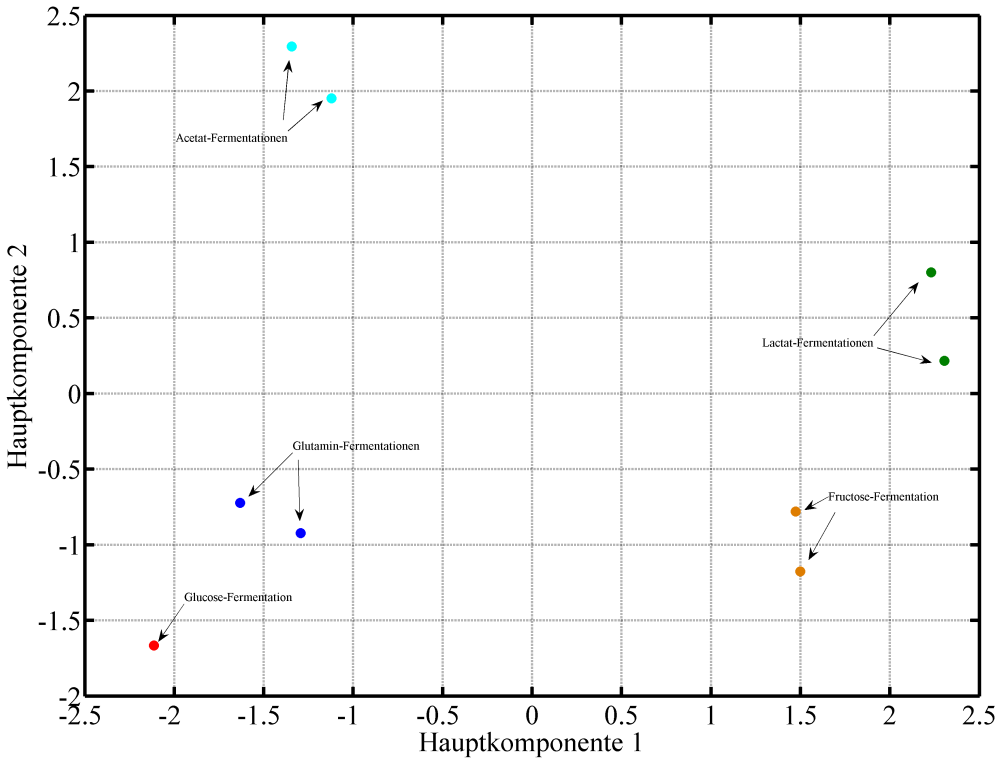

Abbildung 7.25: Hauptkomponentenanalyse basierend auf den gemeinsamen paarweisen Prozessähnlichkeiten, abgegriffen durch die Pearson'sche Korrelation. Verwendete Daten: Glucose-, Fructose-, Acetat-, Lactat- und Glutamin-Fermentationen. Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung. Anzahl zugrunde liegender paarweiser Metabolitkombinationen: 854.

Im Gegensatz zur Hauptkomponentendarstellung basierend auf den vorverarbeiteten Konzentrationsdaten (vergleiche Kapitel 7.1.5) wurde in diesem Fall ein vergleichbar geringer Ausgangsdatensatz verwendet. So finden hierbei nicht al-

le Metaboliten, sondern nur diejenigen die gleichermaßen in allen Experimenten vorkommen und zwischen denen auch laut theoretischer Betrachtung ein metabolischer Pfad existiert, Berücksichting. Ferner ist anzumerken, dass durch den prozessbezogenen Charakter der Deskriptoren die zeitliche Dimension in dem betrachteten Datensatz nicht mehr vorhanden ist. Letztere Eigenschaft der Deskriptoren prädestiniert sie besonders für die Identifikation von Biomarkern im Kontext diagnostischer Fragestellungen basierend auf Metabolomdaten.

7.5.1 Substratinduzierte Unterschiede im Stoffwechsel

Um in einem zweiten Schritt herauszufinden, welche der Metabolitpaarungen (als Variablen betrachtet) sich hinsichtlich ihrer paarweisen Prozessähnlichkeit am stärksten zwischen den Fermentationsexperimenten unterscheiden, wurde der bestehende Datensatz mit Hilfe von Verfahren der Merkmalsselektion untersucht, wobei in diesem Zusammenhang der BestFirst-Algorithmus genutzt wurde (Pearl, 1984). Die Merkmalsselektion ergab, dass eine Untergruppe von 26 der 854 Variablen (also Paarungen von Metaboliten) bereits vollkommen ausreicht, um die Substratgruppen unter Berücksichtigung der Wiederholungsexperimente zu trennen.

Die nachfolgende Tabelle 7.9 listet die ausgewählten 26 Metabolitkombinationen mit den stärksten Trenneigenschaften nach ihrer Wichtigkeit und mit ihrem entsprechenden Reaktionsabstand auf. Für den Reaktionsabstand wurden exemplarisch die Ergebnisse der VGL2-Modellierung unter dem CUBIC-Mapping gewählt, da diese (wie ausführlich besprochen) generell die kürzesten Abstände verglichen mit anderen Modellierungen und dem KEGG-Mapping lieferte.

Die Tabelle zeigt, dass die Metabolitpaarungen mit den besten Trenneigenschaften hinsichtlich des verwendeteten Ausgangssubstrates, nicht unmittelbar im metabolischen Netzwerk benachbart sind. Im geringsten Fall sind die betreffenden Metaboliten 3 Reaktionsschritte voneinander entfernt, der größte Abstand beträgt 9 Reaktionsschritte. Die Erklärung dieser Auffälligkeiten liegt außerhalb des Betrachtungssystemes dieser Arbeit und muss in einem biochemischen Kontext tiefergehend untersucht werden. Festgestellt werden kann jedoch, dass sich in dieser Liste mit (S)-Malate (C00117), Succinate (C00042) und Citrate (C00158) insgesamt drei Metaboliten des Zitratzyklus - welcher bei allen Fermentationen aktiv sein muss - befinden. Die genannten Metaboliten finden sich in der Tabelle in

Tabelle 7.9: Tabellarische Übersicht der durch Merkmalsselektion ausgewählten Metabolitpaarungen mit den besten Trenneigenschaften. Verwendeter Deskriptor: Pearson'sche-Korrelation. Angabe des zugehörigen Reaktionsabstandes aus der VGL2-Modellierung unter Verwendung des CUBIC-Mappings.

Merkmal-Nr.	Metabolitpaarung	Abstand
1	(S)-Malate vs. L-Proline	5
2	(S)-Malate vs. L-Glutamate	3
3	L-Isoleucine vs. Uridine	6
4	(S)-Lactate vs. L-Alanine	5
5	L-Proline vs. Glycine	4
6	L-Aspartate vs. alpha,alpha-Trehalose	5
7	L-Isoleucine vs. L-Threonine	7
8	L-Isoleucine vs. (S)-Malate	4
9	(S)-Malate vs. L-Homoserine	4
10	beta-D-Fructose 6-phosphate vs. L-Lysine	5
11	D-Glycerate vs. L-Lysine	6
12	Succinate vs. L-Lysine	5
13	Glycolate vs. L-Alanine	7
14	beta-D-Fructose 6-phosphate vs. L-Alanine	8
15	L-Valine vs. beta-D-Fructose 6-phosphate	4
16	(S)-Lactate vs. L-Lysine	6
17	Glycolate vs. alpha,alpha-Trehalose	9
18	L-Homocysteine vs. beta-D-Glucose	9
19	beta-D-Glucose vs. Uridine	7
20	Glycolate vs. D-Ribose	8
21	Glycolate vs. (S)-Lactate	4
22	L-Glutamate vs. Glycine	3
23	D-Glycerate vs. L-Glutamate	4
24	(S)-Malate vs. L-Lysine	5
25	Citrate vs. D-Xylulose 5-phosphate	5
26	Citrate vs. D-Ribose 5-phosphate	5

Kombinationen mit Aminosäuren wie L-Proline (C00148), L-Glutamate (C00025), L-Isoleucine (C00407), L-Homoserine (C00263) oder L-Lysine (C00047), die Endprodukte des Stoffwechsel darstellen. Diese Kombinationen lassen den Schluss zu, dass die Aminosäuren in Abhängigkeit vom Ausgangssubstrat in unterschiedlichen Konzentrationsverhältnissen hergestellt werden. Die unterschiedlich stark ausgeprägte Produktion von Aminosäuren konnte bereits mehrfach in Studien, welche die biotechnologische Verwertbarkeit von coryneformen Bakterien untersuchten, festgestellt werden (Hermann, 2003).

Abbildung 7.26 zeigt die 26 ausgewählten Variablen mit ihren zugehörigen Prozessähnlichkeiten aus allen Fermentationsexperimente in einer Heatmap-Darstellung. In ihr sind die signifikanten Unterschiede im Stoffwechsel von *C. glutamicum* unter den untersuchten Ausgangssubstraten in kondensierter Form dargestellt. In Zeilenform sind die 26 Variablen (Metabolitkombinationen) mit den besten Trenneigenschaften angeordnet. Die paarweisen Prozessähnlichkeiten zwischen den Paarungen sind als farbliche Kästchen kodiert. In roten Farbtönen sind die positiven Korrelationen dargestellt, in grün die negativen Korrelationen. Je stärker der Rot-Ton, desto näher liegt die Korrelation am Wert 1 und je stärker der Grün-Ton, desto näher liegt die Korrelation an -1. Korrelationen um den Wert Null sind in dunklen Farbtönen gekennzeichnet. Die Grafik zeigt, dass die einzelnen Experimente fehlerlos den Substratgruppen zugeordnet werden können. Des weiteren ist zu erkennen, dass sich die Prozessähnlichkeit - abgegriffen durch die Pearson'sche Korrelation - zwischen gewissen Metabolitpaarungen stärker ähnelt.

So sind beispielsweise vom Metaboliten Citrate (C00158) ausgehend, Kombinationen zu D-Xylulose 5-phosphate (C00231) und D-Ribose 5-phosphate (C00117), beides wichtige Metaboliten des Pentose-Phosphat-Weges, festzustellen. Dies lässt (unter der Voraussetzung, dass der Zitratzyklus immer aktiv sein muss) den interpretatorischen Schluss zu, dass der Pentose-Phosphat-Weg vermutlich nicht in allen Fermentationen gleichermaßen stark frequentiert ist. Aus den Untersuchungen der enzymkodierenden Gentranskripte unter bestimmten Wachstumsbedingungen (vergleiche 7.4.3.1) konnte bereits abgeleitet werden, dass der Pentose-Phosphat-Weg im Falle der Acetat-Fütterung deutlich heruntergeregelt ist. Schaut man sich die beiden letzten Zeilen der Abbildung 7.26 an, könnte dies ein Hinweis darauf sein, dass eine starke Herunterregelung der Aktivität des Pentose-Phosphat-Weges auch für Fütterungsbedingungen unter Lactat gilt. Für das Wachstum unter Fructose und Glutamin finden sich hingegen positive Korrelationen, welche durch röt-

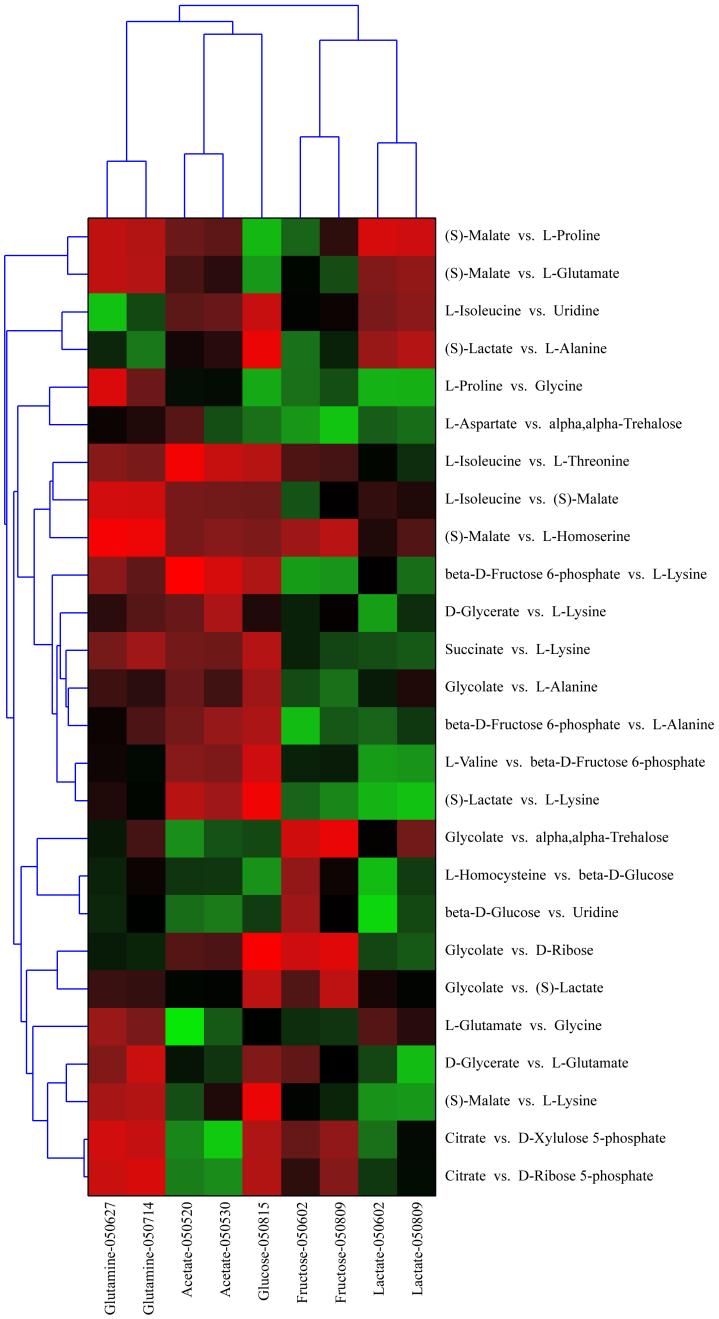

Abbildung 7.26: Heatmapdarstellung der durch Merkmalsselektion ausgewählten Metabolitpaarungen höchster Trenneigenschaft. Zugrundeliegenden sind die in allen Fermentationen gemeinsamen Metabolitpaarungen (n=854). Berechneter Deskriptor: Pearson-Korrelation. Verwendete Daten: Glucose-, Fructose-, Acetat-, Lactat- und Glutamin-Fermentationen. Datenvorverarbeitung: adaptive Fehlerkorrektur, Ausreißerkorrektur. Datenskalierung: Logarithmierung und Medianzentrierung.

liche Farbtöne dargestellt sind. Ob hier das umgekehrte Phänomen gilt, kann nur unter Heranziehung zusätzlicher Daten (beispielsweise Transkriptomuntersuchungen bei Fütterung mit Fructose und Glutamin unter vergleichbaren Randbedingungen) beantwortet werden.

7.5.2 Substratinvariante Merkmale im Stoffwechsel

Im umgekehrten Fall wurde auch untersucht, welche Metabolitpaarungen sich hinsichtlich ihrer Prozessähnlichkeit konservativ verhalten, das heißt sich folglich nur wenig über die Fermentationsexperimente hinweg ändern. Die Tabelle 7.10 zeigt absteigend angeordnet die 20 Metabolitpaare geringster Veränderung. Zur Erstellung dieser Tabelle wurde exemplarisch das Maß der Pearson'schen Korrelation als Deskriptor verwendet. Es konnte festgestellt werden, dass bei manchen der betrachteten Paarungen mindestens einer, manchmal auch beide Partner aus dem Zitratzyklus stammen. Im Gegensatz zu Tabelle 7.9 können hierbei auch Paarungen gefunden werden, die sich in direkter Nachbarschaft im metabolischen Netzwerk befinden. So verhält sich beispielsweise die Prozessähnlichkeit der Metaboliten Succinate (C00042) und Fumarate (C00122) invariant in Bezug auf die verwendeten Ausgangssubstrate. Für diese Paarung konnte bereits in der integrativen Analyse unter Heranziehung der Transkriptomdaten festgestellt werden, dass sich trotz einer signifikanten Exprimierung des Enzymgens die Prozessähnlichkeit nur geringfügig verändert. Eine weitere Metabolitpaarung, die sich in direkter Nachbarschaft befindet und sich substratinvariant verhält, ist die Paarung zwischen L-Aspartate (C00049) und beta-Alanine (C00099). Katalysierendes Enzym ist Aspartate 1-Decarboxylase (EC 4.1.1.11), welche L-Aspartate irreversibel unter Freisetzung von CO_2 in beta-Alanine umsetzt. Für beide, sich konservativ verhaltenden Metabolitpaarungen in direkter Nachbarschaft könnte gelten, dass sie besonders wichtig für den Stoffwechsel sind und daher invariant auf unterschiedliche Umweltbedingungen reagieren.

Zusammenfassung

Es zeigt sich deutlich, dass die Deskriptoren zur Bestimmung der paarweisen Prozessähnlichkeit von Metabolitzeitreihen die substratinduzierten Unterschiede im Stoffwechsel von *C. glutamicum* gut abgreifen können und dass das Potenzial für diskriminatorische Anwendungen besitzen. Die Untersuchung der gleichermaßen

Tabelle 7.10: Tabellarische Übersicht der 20 Variablen mit geringster Variation in ihrer Prozessähnlichkeit unter Berücksichtigung aller Fermentationen hinweg. Verwendeter Deskriptor: Pearson'sche Korrelation unter Verwendung der Fisher-Transformation. Angabe des zugehörigen Reaktionsabstandes aus der VGL2-Modellierung unter Verwendung des CUBIC-Mappings.

Merkmal-Nr.	Metabolitpaarung	Abstand
1	D-Glucono-1 5-lactone vs. D-Ribose 5-phosphate	6
2	L-Alanine vs. Mannitol	11
3	L-Alanine vs. Citrate	3
4	D-Xylulose 5-phosphate vs. D-Ribose 5-phosphate	2
5	Succinate vs. Fumarate	1
6	L-Homocysteine vs. Fumarate	3
7	Glycine vs. L-Lysine	7
8	Succinate vs. L-Homoserine	4
9	Glycerol vs. L-Phenylalanine	6
10	L-Threonine vs. (S)-Malate	4
11	Glycolate vs. beta-D-Fructose 6-phosphate	7
12	D-Glycerate vs. 2-Oxoglutarate	3
13	L-Isoleucine vs. L-Tyrosine	2
14	N-Acetyl-L-glutamate vs. Citrate	3
15	L-Isoleucine vs. L-Valine	2
16	L-Homoserine vs. L-Homocysteine	2
17	L-Aspartate vs. beta-Alanine	1
18	L-Serine vs. D-Ribose	6
19	L-Alanine vs. 2-Oxoglutarate	5
20	L-Aspartate vs. L-Lysine	3

in allen Fermentationsexperimenten vorhandenen Metabolitpaarungen und ihrer Prozessähnlichkeiten ergab, dass der Stoffwechsel nicht gleichartig stark auf die unterschiedliche Umweltbedingungen reagiert. Es existieren einerseits Metabolitpaarungen, deren Prozessähnlichkeit sich deutlich in Abhängigkeit des vorhandenen Ausgangssubstrates verändern und solche, die sich annähernd gleichartig hinsichtlich ihrer Prozessähnlichkeit verhalten. Die Interpretation der ermittelten Paarungen größter und geringster Veränderung ist nicht trivial, da die Paarungen sich nicht immer in direkter Nachbarschaft des metabolischen Netzwerkes befinden. Die Tatsache, dass sowohl stark variable als auch konservative Paarungen gefunden wurden, könnte ein Hinweis darauf sein, dass Teile des Stoffwechsels variabel gesteuert werden können (oder müssen), während wieder andere Teile robust gegen Umwelteinflüsse immer gleichartig funktionieren können (beziehungsweise müssen).

Unbedingt anzumerken bleibt in diesem Zusammenhang, dass die durchgeführte Suche von Unterschieden und Gemeinsamkeiten im Stoffwechsel von *C. glutamicum* selbstverständlich von der Stichprobengröße der betrachteten Experimente abhängig ist. Besonders die Ergebnisse aus Kapitel 7.5 stellen daher erste Hinweise dar; eine Vergrößerung des Stichprobenumfanges ist in weiterführenden Untersuchungen anzustreben und würde die Robustheit der Aussagen deutlich erhöhen.

8 Diskussion

Metabolismus ist eine Kombination komplexer und dynamischer Prozesse, die auf molekularer Ebene stattfinden. Um diesen prozessorientierten Charakter zu berücksichtigen, beschäftigt sich diese Arbeit intensiv mit der Betrachtung der Zeitreiheneigenschaften experimentell erfasster Metabolitkonzentrationen. Dies unterscheidet diese Arbeit grundlegend von anderen Studien, welche den mikrobiellen Metabolismus oft nur in Einzelmessungen und bestenfalls in Replikaten betrachten. Zur Betrachtung der Zeitreiheneigenschaften wurden eigene Deskriptoren entwickelt, da sich traditionelle Verfahren der Zeitreihenanalyse aufgrund der vergleichsweise geringen Anzahl von Datenpunkten als nicht applikabel erwiesen.

In Vorversuchen zeigte sich, dass für eine optimale Analyse der betrachteten Metabolitzeitreihen eine Vorverarbeitung der Daten unerlässlich ist. Hierbei wurde die Vorverarbeitung in zwei grundlegend unterschiedliche Teilbereiche unterteilt. Zuerst wurden im Rahmen dieser Arbeit Ansätze zur adaptiven Korrektur der Konzentrationsdaten unter Berücksichtig von Ausreißern und fehlenden Werten (infolge zu kleiner Peakflächen) entwickelt. Nachfolgend wurden die Metabolitzeitreihen mit Hilfe geeigneter Transformations- und Skalierungsverfahren vergleichbar gemacht. Der Grund für dieses Vorgehen besteht darin, dass sich Metaboliten von ihren Konzentrationen her um mehrere Größenordnungen unterscheiden können und folglich ohne Datentransformation jene Metaboliten höchster Konzentration die Analyse dominieren. Aus den Vorversuchen wurde deutlich, dass ein paarweiser Vergleich zwischen Metabolitzeitreihen nur nach einer Transformation sinnvoll durchgeführt werden kann. Zum Finden der geeigneten Vorverarbeitungsstrategie wurden zahlreiche Voruntersuchungen im Batch-Verfahren durchgeführt. Es zeigte sich, dass eine Kombination von Transformationsverfahren wie beispielsweise der Logarithmierung und Skalierungsverfahren wie der Medianzentrierung am ehesten dazu geeignet waren, Prozessstrukturen aus den Daten herauszuarbeiten.

Die Fragestellung, ob sich in den Konzentrationsdaten Strukturen finden lassen, wurde mit unüberwachten Verfahren der multivariaten Statistik auf optimal vorverarbeiteten Konzentrationszeitreihen bearbeitet. Es konnte demonstriert werden, dass interessante Strukturen in den Daten zu finden sind. Clustert man beispielsweise die experimentellen Daten über alle Metaboliten hinweg, so wird sichtbar, dass sich der gesamte Stoffwechsel entlang der Fermentationsexperimente tiefgreifend verändert und dass deutlich unterschiedliche Phasen entlang der Zeitachse festzustellen sind (Kapitel 7.1). Diese durch die Clusteranalyse gelieferte Einteilung entspricht nahezu exakt den physiologischen Wachstumsphasen, wie sie beispielsweise auch von der Betrachtung der optischen Dichte (Abbildung 4.4) oder anderen Ansätzen her abgeleitet werden kann. Während die Lag-Phase eher durch Anpassungsvorgänge an das Nährmedium gekennzeichnet ist, steht beispielsweise in der exponentiellen Wachstumsphase die vornehmliche Produktion von Biomasse für das Zellwachstum im Vordergrund. In der stationären Wachstumsphase ist das Ausgangssubstrat aufgebraucht und es finden unter Umständen autolytische Prozesse statt. Aufgrund dieser Ergebnisse kann davon ausgegangen werden, dass sich diese übergeordneten Grundmotive, denen der Stoffwechsel von *C. glutamicum* folgt, bis in die experimentellen Daten durchpausen.

Clustert man die experimentellen Daten in entgegengesetzter Dimension, das heisst entlang der Zeitachse, so sind ebenfalls deutliche Gruppen festzustellen. Hier gruppieren sich jene Metaboliten, welche ein ähnliches Prozessverhaltens im Verlauf des Fermentationsexperimentes aufweisen. Diesen Effekt der Clusterbildung findet man am deutlichsten in der Glucose-Fermentation (Kapitel 7.1.4.2), er ist jedoch auch bei allen anderen Fermentationen feststellbar. Drei grundlegende Typen temporalen Konzentrationsverhaltens konnten hierbei identifiziert werden. Der erste Typus zeichnet sich durch eine stetige Verringerung der Konzentration im zeitlichen Verlauf des Fermentationsexperimentes aus. Der zweite und dritte Typus akkumulieren höhere Konzentrationen, wobei der zweite Typus nur bis zum Ende der exponentiellen Phase ansteigt und dann wieder abfällt, während sich der dritte Typus durch einen stetigen Anstieg - bis in die stationäre Wachstumsphase hinein - charakterisieren lässt. Eine solch deutliche Einteilung experimentell erfasster Metabolitkonzentrationen in Gruppen unterschiedlichen Prozessverhaltens konnte nach aktuellem Wissenstand bisher in keiner anderen Studie beobachtet werden.

Für die Zuordnung der Metaboliten in ihre jeweiligen Cluster zeigten sich zahl-

reiche Interpretationsmöglichkeiten auf. So finden sich beispielsweise jene Metaboliten, die gewissermaßen als Vorläufermetabolite direkt aus dem Ausgangssubstrat umgewandelt werden, im Typus, welcher sich durch kontinuierliches „Leerlaufen“ kennzeichnet. Nahezu alle Aminosäuren, welche eher als Endprodukte des Stoffwechsels angesehen werden können, befinden sich in jener Gruppe, welche sich durch stetige Akkumulation kennzeichnet. Diese Einteilung, welche ohne topologische Zusatzinformationen gewonnen werden konnte, zeigt deutlich, dass Metaboliten unterschiedliche Positionen und Funktionen im Stoffwechsel haben müssen. Vorläufermetabolite müssen in ihrer Konzentration abnehmen, damit überhaupt erst andere Metaboliten wie beispielsweise Aminosäuren, im größerem Umfang hergestellt werden können. Es bleibt anzumerken, dass die oben genannten Auffälligkeiten bei der unüberwachten Datenstrukturanalyse erst nach Durchführung einer geeigneten Datenvorverarbeitungsstrategie, welche sowohl eine adaptive Fehlerkorrektur (Kapitel 5.1.1.3 und 5.1.1.4), als auch Verfahren zur Datentransformation (Kapitel 5.1.1.5) beinhaltete, sichtbar wurden.

Wie auch belegt werden konnte, verhalten sich manche Metaboliten in ihrem Prozessverhalten zueinander ähnlicher als andere. Da eine hohe Prozessähnlichkeit ein Hinweis auf eine mögliche enzymatische Verknüpfung der beteiligten Metaboliten sein kann (dieser Ansatz findet bei der *de novo* Rekonstruktion von metabolischen Pfaden aus Korrelationsdaten Verwendung) wurde im weiteren Verlauf der Arbeit beleuchtet, welche Metaboliten eine hohe Prozessähnlichkeit zueinander aufweisen, und was die Ursache hierfür sein kann. Um diese Fragestellung zu klären, wurde das metabolische Netzwerk von *C. glutamicum* auf Basis einer Genomannotation (Kapitel 4.3.1) rechnergestützt rekonstruiert. Hierbei wurde schrittweise vorgegangen und Informationen aus externen Datenbanken berücksichtigt. Sämtliche Zusatzinformationen wurden manuell anhand von Expertenwissen überprüft. Letztendlich wurden nach mehreren Arbeitsschritten zwei theoretische Reaktionsnetzwerke für *C. glutamicum* erzeugt (Vergleiche hierzu insbesondere 4.3.2). Die beiden Reaktionsnetzwerke unterscheiden sich dergestalt, dass ersteres sich strikt an die Erkenntnisse aus der Genomannotation hält, während das andere zusätzliche hypothetische Information über potenzielle Lückenfüller im Stoffwechsel enthält.

Basierend auf den erzeugten Netzwerken wurden weitere Analysen durchgeführt. Hierbei wurden die Netzwerke mit Hilfe graphentheoretischer Ansätze analysiert. Ziel dieses Vorgehens war es einerseits, biochemisch plausible Stoffwech-

selwege zu finden und andererseits weitere beschreibende Größen abzuleiten. Zu diesen topologischen Größen, die den Metaboliten innerhalb seines Netzwerkes charakterisieren, gehören beispielsweise der Verknüpfungsgrad eines Metaboliten zu einen Nachbarn oder der kürzeste Reaktionsabstand zweier Metaboliten zueinander. Sämtliche Netzwerkmodellierungen wurden unter Berücksichtigung der Problematik von Seitenmetaboliten durchgeführt.

Die experimentellen als auch die theoretischen Informationen wurden in einer integrativen Analyse weitergehend untersucht. Der Grundgedanke dieses Ansatzes besteht darin, Informationen aus unterschiedlichen Quellen miteinander zu verbinden und nach Wechselwirkungen zu suchen. Im Falle dieser Arbeit wurden unter anderem aus experimentellen Zeitreihen ableitbare Deskriptoren mit theoretischen Beschreibungen derselben in Zusammenschau analysiert. Dieses Vorgehen ist zulässig, da das betrachtete Untersuchungsobjekt jeweils identisch ist. Ziel der Analyse war es zu untersuchen, ob sich die in den Zeitreihen gefundenen Strukturen unter Zuhilfenahme der theoretischen Informationen weitergehend klären lassen.

Bringt man nun beispielsweise den theoretischen Reaktionsabstand im theoretischen Netzwerk mit der Prozessähnlichkeit der betrachteten Metaboliten in Verbindung zeigt sich, dass Paare signifikant hoher Prozessähnlichkeit sowohl in direkter Nachbarschaft als auch im metabolischen Netzwerk weit voneinander entfernt auftreten können (Kapitel 7.4.2.1). Es kann folglich nicht festgestellt werden, dass ein übergeordneter Zusammenhang zwischen der Prozessähnlichkeit und dem Reaktionsabstand zweier Metaboliten vorhanden ist. Jedoch bedürfen die gefundenen hohen Prozessähnlichkeiten einer Interpretation. Hohe Prozessähnlichkeit in direkter Nachbarschaft ist ein vergleichsweise oft zu findendes Phänomen. Es tritt beispielsweise dann auf, wenn Metaboliten infolge enzymatischer Aktivität direkt ineinander überführt werden, oder anders ausgedrückt: Wenn der metabolische Fluss über die beiden Metaboliten führt. Für das Auftreten hoher Prozessähnlichkeiten bei großem Reaktionsabstand konnte eine Arbeitshypothese angeführt werden. So weisen häufig Paarungen von Aminosäuren hohe Prozessähnlichkeiten auf, obwohl sie nur durch eine große Anzahl von Reaktionsschritten ineinander umgesetzt werden können. Eine direkte enzymatische Regulation scheidet folglich nach aktuellem Wissensstand aus. Als Erklärung für die hohe Prozessähnlichkeit ist vielleicht die parallele, bedarfsgesteuerte Produktion der Aminosäuren anzuführen. In diesem Fall bedeutet dies, dass die betreffenden

Metaboliten gleichermaßen von einem übergeordneten Prozess reguliert werden, welcher sich außerhalb der Betrachtungsweise dieser Arbeit befindet. Aufgrund der hohen Prozessähnlichkeiten ist ferner anzunehmen, dass die parallele Produktion von Aminosäuren relativ isoliert und frei von anderen Einflüssen (bzw. Interaktion mit zahlreichen anderen Reaktionspartnern) stattfindet.

Warum jedoch hohe Prozessähnlichkeiten zwischen Metaboliten auftreten, konnte anhand der abgeleiteten Topologiedeskriptoren weiter eingeschränkt werden. So treten signifikant hohe Prozessähnlichkeiten nur dann auf, wenn die beiden beteiligten Metaboliten ihrerseits jeweils wenige Nachbarn haben (Kapitel 7.4.2.4) und sich hinsichtlich ihrer Konzentrationen nicht allzu stark voneinander unterscheiden (Kapitel 7.4.2.3). Dies ist insofern bemerkenswert, da diese Auffälligkeiten in Unabhängigkeit von der Entfernung der Metaboliten in nahezu allen Fermentationen deutlich festgestellt werden können. Im Umkehrschluss bedeutet dies, dass eine hohe Prozessähnlichkeit zwischen Metaboliten dann nicht auftreten kann, wenn mindestens einer der Metaboliten stark verknüpft ist und somit wie beispielsweise Pyruvate eher die Funktion eines „Metabolit-Hubs“ einnimmt. Auch dies konnte durch Untersuchung der gemittelten Prozessähnlichkeit von Metaboliten zu ihren Nachbarn im metabolischen Netzwerk bewiesen werden.

Neben dem paarweisen Vergleich von Metaboliten, wurde auch eine metabolitzentrische Betrachtung durchgeführt. Dies bedeutet, dass die Eigenschaft der Konzentrationszeitreihe als solche in Zusammenschau mit theoretischen Netzwerktopologien untersucht wurde. Hierbei konnte dargestellt werden, dass Metaboliten mit einem hohen Verknüpfungsgrad zu anderen Metaboliten nur in vergleichsweise hohen Konzentrationen vorkommen (Kapitel 7.4.1.1); schwach verknüpfte Metaboliten kommen ausschließlich in niedrigen Konzentrationen vor. Aus diesen Ergebnissen kann abgeleitet werden, dass die Konzentration, in der ein Metabolit gemessen wird, von dessen Position im metabolischen Netzwerk abhängig ist. Ein Metabolit, welcher beispielsweise im Zentralstoffwechsel mit vielen anderen Metaboliten durch Reaktionen verbunden ist und somit Teil vieler metabolischer Flüsse ist, muss *per se* in vergleichbar höheren Konzentrationen vorliegen. Ein Metabolit, der nur sehr wenige Nachbarn besitzt und beispielsweise an der Peripherie des Netzwerkes anzusiedeln ist, kann keine beliebig hohe Konzentration annehmen, da diese von der Konzentration seiner wenigen Nachbarn, sowie der enzymatischen Aktivität stark limitiert ist. Auch für die Sensitivität der Zeitreihe kann in Zusammenschau mit dem theoretischen Verknüpfungsgrad eine Auffällig-

keit festgestellt werden (Kapitel 7.4.1.4). Hochverknüpfte Metaboliten weisen in der Regel eine geringe Schwankungsbreite auf, wohingegen schwach verknüpfte Metaboliten die höchsten Schwankungen in der Zeitreihe besitzen. Dies kann dergestalt erklärt werden, dass hochverknüpfte Metaboliten an vielen Reaktionen partizipieren, welche sich unter Umständen im Konzentrationsverlauf gegenseitig aufheben. Besitzt ein Metabolit nur wenige Nachbarn, ist sein Konzentrationsverlauf in hohem Maße von der Aktivität der katalysierenden Enzyme abhängig.

Zusammengefasst kann gesagt werden, dass die Metabolitprofile unter Zuhilfenahme der theoretischen Netzwerktopologien deutlich besser interpretiert werden können. Vor allem die Betrachtung der Nachbarschaftsverhältnisse von Metaboliten hilft zu verstehen, warum sich Metaboliten in ihrem temporalen Konzentrationsverlauf so verhalten, wie sie es tun. Die Betrachtung der Nachbarschaftsverhältnisse erlaubte es auch, das Auftreten signifikant hoher Prozessähnlichkeiten unabhängig vom Reaktionsabstand einzugrenzen. Für die hohen Prozessähnlichkeiten von Metaboliten unter großem Reaktionsabstand liefert diese Arbeit mit der bedarfsgesteuerten Parallelproduktion einen möglichen Erklärungsansatz. Weitere Untersuchungen sollten jedoch folgen, um die Steuerungsmechanismen der parallelen Produktion von Aminosäuren stärker zu beleuchten.

Die theoretischen Netzwerke und die daraus abgeleiteten Topologien jedoch sind statische Informationen. Da sie gewissermaßen nur die *Möglichkeit* einer Umsetzung zwischen Metaboliten repräsentieren, lag hier ein Ansatzpunkt für die Integration von Zusatzinformation. Diese wurde aus Untersuchungen der Gentranskripte, wie die von anderen Forschergruppen an *C. glutamicum* durchgeführt worden sind, extrahiert. Die Untersuchung der Gentranskripte lag ebenfalls für verschiedene Fermentationsbedingungen vor und lieferte Ansatzpunkte darüber, welche enzymkodierenden Gene unter welchen Wachstumsbedingungen signifikant exprimiert sind. Für Wachstumsbedingungen bei Fütterung mit Glucose und Acetat wurde die Untersuchung detailliert durchgeführt, denn hier lagen Information über die transkriptionelle Aktivität aus drei unabhängigen Studien vor (Gerstmeir et al., 2003; Hayashi et al., 2002; Muffler et al., 2002). Im Fokus der erweiterten integrativen Analyse lagen diesbezüglich primär die Prozessähnlichkeiten direkt im theoretischen Netzwerk benachbarter Metaboliten, sowie die Exprimierung des korrespondierenden enzymkodierenden Gens. Interessanterweise konnte für mehrere benachbarte Metaboliten des Zentralstoffwechsels (Kapitel 7.4.3) eine deutlich höhere Prozessähnlichkeit dann nachgewiesen werden, wenn bei den entspre-

chenden Fermentationsbedingungen das betreffende Enzymgen stärker exprimiert war. Der umgekehrte Fall konnte nicht beobachtet werden. Dieser Zusammenhang von gesteigerter Expression des Enzymgens und erhöhter Prozessähnlichkeit im Konzentrationsverlauf benachbarter Metaboliten konnte bisher in keiner wissenschaftliche Arbeit nachgewiesen werden. Er kann dergestalt interpretiert werden, dass die hohe Prozessähnlichkeit über die Menge des katalysierenden Enzyms sichergestellt wird. Enzyme fungieren als Biokatalysatoren und ermöglichen Reaktionen, welche ohne ihre Beteiligung nicht ablaufen würden, hohe Prozessähnlichkeit zweier benachbarter Metaboliten kann hierzu ein entscheidender Hinweis sein. Diese Ergebnisse zeigen eindrucksvoll, dass sich im Transkriptom gesteuerte Prozesse durchaus bis auf die Ebene des Metaboloms durchpausen können. Die Wichtigkeit, biologische System als Ganzes zu betrachten und nicht einzelne Informationsebenen (Genom, Transkriptom, Proteom, Metabolom) ausschließlich individuell zu analysieren, wird hierdurch untermauert.

In einer abschließenden Analyse wurde untersucht, ob die entwickelten Deskriptoren zur Charakterisierung der Prozessähnlichkeit auf den Konzentrationszeitreihen auch für diskriminatorische Zwecke Verwendung finden können (Kapitel 7.5). Hierzu wurden 854 Metabolitpaarungen, welche gemeinsam in allen Fermentationsexperimenten vorkamen, in einer gemeinsamen Analyse untersucht. Die Analyse ergab, dass die durch verschiedene Ausgangssubstrate induzierten Unterschiede im Stoffwechsel gut durch die Deskriptoren abgegriffen werden können. Eine Hauptkomponentenanalyse zeigte, dass die jeweiligen Substratgruppen deutlich voneinander getrennt werden können (Abbildung 7.25). Weiterhin wurde deutlich, dass bereits ein Subset aus wenigen ausgewählten Metabolitpaarungen die Unterschiede im Stoffwechsel hinreichend beschreiben kann (Abbildung 7.26). Es konnten ferner auch jene Metabolitpaarungen ermittelt werden, welche sich hinsichtlich ihrer Prozessähnlichkeit über die Fermentationen hinweg am wenigsten verändern (Tabelle 7.10). Zusammengefasst bedeutet dies, dass prozessorientierte Deskriptoren zur Ähnlichkeitsberechnung auf Konzentrationszeitreihen von Metaboliten das Potenzial zur Detektion von signifikanten Unterschieden als auch von konservativem Verhalten im Stoffwechsel von Organismen besitzen. Die Detektion substratinduzierter als auch substratinvarianter Merkmale im Stoffwechsel von *C. glutamicum* ermöglicht es, weitere wertvolle Erkenntnisse für die biotechnologische Verwertung des Bakteriums zu erlangen und sollte daher weiterverfolgt werden.

Aus sämtlichen Ergebnissen resultiert die Forderung, dass für das bessere Verständnis komplexer biologischer Systeme in zukünftigen Studien Daten aus unterschiedlichen „Omics“-Bereichen einer integrativen und prozessbezogenen Analyse unterzogen werden sollten. Die integrative Analyse beispielsweise von Transkriptom, Proteom und Metabolom könnte Zusammenhänge offenbaren, welche sich aus der losgelösten Betrachtung der einzelnen Interaktionsebenen bei weitem nicht ergeben. Oder vereinfacht ausgedrückt: die Betrachtung des Ganzen wird weit mehr Erkenntnisse liefern als die Summe seiner Bestandteile.

Ferner zeigte die Arbeit auf, dass anstelle von Punktmessungen engmaschige Zeitreihen der Metabolitkonzentration erfasst werden sollten, um den prozessorientierten Charakter des Stoffwechsels tiefergehend zu beleuchten. In Zukunft kann davon ausgegangen werden, dass die fortschreitende Entwicklung in der instrumentellen Analytik diese Hochdurchsatzanalytik in kurzen zeitlichen Abständen auch tatsächlich erlaubt. Ferner ist im Hinblick auf die Metabolomanalyse davon auszugehen, dass die Anzahl detektier- und indentifizierbarer Metaboliten ebenfalls ansteigt und somit eine komplexere Zusammenschau ermöglicht. Für die Analyse der heterogenen Daten aus unterschiedlichen Quellen sind sind geeignete Werkzeuge, Datenstandards und Nomenklaturen von Nöten, welche die gemeinsame Betrachtung und Analyse in einer geeigneten Art und Weise ermöglichen. Ein erster, viel versprechender Ansatz in der Schaffung von Standards bei der theoretischen Repräsentation biologischer Systeme ist hierbei in der MIRIAM-Initiative von Le Novère et al., 2005 zu sehen, welche sich in der wissenschaftlichen Gemeinde zunehmend etabliert.

Es bleibt abzuwarten, welche spannenden regulatorischen Erkenntnisse sich aus dieser integrativen Betrachtungsweise entwickeln werden. Ein tiefergehendes Verständnis biologischer Systeme könnte zahlreiche Möglichkeiten eröffnen - beispielsweise für die Entwicklung neuer Medikamente, die Detektion von spezifischen Biomarkern oder die Entwicklung individualisierter Therapieansätze - welche in der bisherigen Betrachtungsweisen versagt bleiben.

9 Zusammenfassung

Ausgewählte, im Rahmen dieser Arbeit erlangte Erkenntnisse werden abschließend an dieser Stelle in Kurzform zusammengefasst:

Unüberwachte statistische Analyse von Zeitreihen der Metabolitkonzentration

- Die physiologischen Wachstumsphasen von *C. glutamicum* können deutlich bei Clusterung der Konzentrationsdaten detektiert werden. Dies ist als Indiz zu werten, dass der Stoffwechsel im zeitlichen Verlauf verschiedenen übergeordneten Themen folgt.

- Bei Clusterung der Konzentrationsdaten in zeitlicher Dimension lassen sich mindestens drei verschiedene Cluster, welche unterschiedliches temporales Prozessverhalten von Metaboliten charakterisierten, feststellen. Erstens: stetige Verringerung der Metabolitkonzentration bis zum Ende der Fermentationsexperimente; zweitens: stetiger Anstieg bis zum Ende der Experimente; sowie drittens ein Konzentrationsverlauf, welcher dem Verlauf der optischen Dichte ähnelt.

- Die Zuordnung von Metaboliten in ihre jeweilige Cluster, zeigt deutlich, dass Metaboliten unterschiedliche Positionen und Funktionen im Stoffwechsel haben müssen. So konnte fermentationsübergreifend beobachtet werden, dass Aminosäuren (als Endprodukte des Stoffwechsels) immer dem gleichen Cluster zugeordnet werden.

- Eine adäquate Datenvorverarbeitung (bestehend aus einer adaptiven Fehlerkorrektur, einer Korrektur für Ausreißer, sowie einer geeigneten Datentransformation) ist für das Finden von Strukturen in den Konzentrationszeitreihen unerlässlich.

Integrative Analyse experimenteller Metabolitzeitreihen mit theoretischen Netzwerktopologien

Metabolitspezifische Betrachtung:

- Metaboliten, die einen vergleichsweise geringen theoretischen Verknüpfungsgrad zu anderen Metaboliten aufweisen, konnten ausschließlich in geringen Konzentrationen detektiert werden. Theoretisch stark verknüpfte Metaboliten weisen hingegen hohe Konzentrationen auf. Stark verknüpfte Metaboliten bei zeitgleich geringer Konzentration konnten nicht festgestellt werden.

- Stark verknüpfte Metaboliten weisen bei allen Fermentationsexperimenten die Konzentrationszeitreihen mit den geringsten Schwankungsbreiten auf. Die höchsten Schwankungsbreiten finden sich hingegen ausnahmslos bei jenen Metaboliten, die einen niedrigen Verknüpfungsgrad aufweisen.

Paarweise Metabolitbetrachtung:

- Für die Mehrzahl der betrachteten Deskriptorenkombinationen aus experimenteller und theoretischer Betrachtung von *C. glutamicum* konnten keine auffälligen Zusammenhänge beobachtet werden.

- Ein genereller, organismenweit gültiger Zusammenhang zwischen der Prozessähnlichkeit zweier Metaboliten (abgegriffen durch Deskriptoren wie das Winkelmaß, Korrelation, etc.) und ihrem korrespondierendem Reaktionsabstand zueinander konnte nicht festgestellt werden.

- Die Betrachtung signifikant korrelierter Metabolitzeitreihen zeigte, dass die betreffenden Metaboliten in ihrem Reaktionsnetzwerk sowohl in direkter Nachbarschaft als auch in großer Entfernung zueinander beobachtet werden konnten.

- Die Bedingungen unter denen eine signifikante Prozessähnlichkeit zweier Metaboliten auftritt, konnte weiter eingeschränkt werden. Signifikant korrelierte Zeitreihen der Metabolitkonzentration traten vorzugsweise dann auf,

wenn jeweils beide Metaboliten einen niedrigen Verknüpfungsgrad zu anderen Metaboliten aufwiesen. Dieser Zusammenhang ist unabhängig vom tatsächlichen Reaktionsabstand der Partner und konnte bei allen Fermentationen nachgewiesen werden.

- Die mittlere Prozessähnlichkeit eines Metaboliten zu seinen (theoretischen) Nachbarn steht in einem deutlichen Zusammenhang zu seinem Verknüpfungsgrad. Hochverknüpfte Metaboliten weisen eine mittlere Korrelation um Null auf, während gering verknüpfte Metaboliten zum Teil einen gemittelten Korrelationskoeffizienten von über r=0,8 erreichen.

- Bei der Fütterung von *C. glutamicum* mit Fructose, Glucose und Glutamin konnte deutlich festgestellt werden, dass die Metabolitpaare, welche sich durch robuste Korrelation auf hohem Siginifikanzniveau auszeichnen, vergleichsweise geringe Konzentrationsunterschiede in der exponentiellen Phase besitzen. Metabolitpaare, welche große Konzentrationsunterschiede aufweisen, treten sehr viel seltener in großer Prozessähnlichkeit auf.

- Die im Rahmen dieser Arbeit entwickelten und verwendeten Deskriptoren zur Beschreibung zeitlichen Prozessverhaltens von Metaboliten, besitzen Potenzial für diskriminatorische Ansätze. So konnten in einer gemeinsamen Analyse aller zur Verfügung stehender Datensätze signifikante, substratinduzierte Unterschiede als auch substratinvariante Merkmale im Stoffwechselverhalten detektiert werden.

Integrative Analyse experimenteller Metabolitzeitreihen mit theoretischen Netzwerktopologien und Transkriptomdaten

- Fermentationsspezifisch konnte für benachbarte Metaboliten im Zentralstoffwechsel von *C. glutamicum* dann eine höhere Prozessähnlichkeit festgestellt werden, wenn das korrespondierende enzymkatalysierende Gen unter den betreffenden Fütterungsbedingungen eine signifikante Expression zeigte. Der umgekehrte Fall konnte in den Daten nicht beobachtet werden.

- Auf Ebene des Transkriptoms gesteuerte Prozesse können sich folglich deutlich in die Zeitreiheneigenschaften betroffener Metaboliten durchpausen. Dies konnte bisher in dieser Form noch nicht beobachtet werden. Auch ohne die direkte analytische Erfassung des Proteoms lassen sich aus den Zeitreiheneigenschaften von Metaboliten erste Rückschlüsse über enzymatische Aktivität treffen.

- Dies demonstriert die für *C.glutamicum* ausgeprägte Anpassungsfähigkeit in Bezug auf unterschiedliche Umweltbedingungen. Durch eine verstärkte Produktion von Enzymen kann der metabolische Fluss an - für den Organismus wichtigen - Stellen im Netzwerk gesteuert werden, was letztendlich die Energieproduktion, die Synthese sämtlicher essentieller Bausteine sowie den Aufbau von Biomasse - kurzum, sein Überleben - ermöglicht.

Nebenergebnisse

- Die im Rahmen dieser Arbeit durchgeführte Genomannotation unter Zuhilfenahme zeitnah analysierter Sequenzdatenbanken wie SwissProt und TrEMBL sowie spezialisierter Datenbanken wie BRENDA resultierte in einem Enzymkatalog für *C. glutamicum*, welcher um ca.10% umfangreicher war, als in den Untersuchungen aus dem Jahre 2003.

- Es konnten einige auffällige Korrelationen zwischen Metaboliten festgestellt werden, zwischen denen nach aktuellem Wissenstand für *C. glutamicum* keine Möglichkeit der Umwandlung existiert. Besonders in zwei Fällen könnte dies als Hinweis für noch nicht entdeckte Enzyme zu werten sein. Für die putativen Enzyme konnte beobachtet werden, dass sie zum Teil in Organismen vorkommen, die als Bodenbakterien ihren natürlichen Lebensraum mit *C. glutamicum* teilen.

- Auch für bis zum jetzigen Zeitpunkt unidentifizierte Substanzen (Unknowns) liefert die paarweise Korrelationsanalyse, wobei hier allerdings alle Datenpunkte der Zeitreihe Verwendung finden, wertvolle Informationen. So konnte für eine hochkorrelierte, nicht identifizierte Substanz in der Zwischenzeit anhand experimenteller Untersuchungen die Vermutung bestätigt werden, dass es sich um ein Derivat eines bereits identifizierten Metaboliten handelt.

- Für die Berechnung metabolischer Pfade in Reaktionsnetzwerken ist eine Korrektur von Wegen über Seitenmetaboliten unerlässlich. Gleiches gilt für die graphische Darstellung metabolischer Netzwerke. Zur Definition von Seitenmetaboliten fehlen allerdings in der noch jungen Wissenschaft der Metabolomforschung allgemeingültige Standards. Vielleicht liefert diese Arbeit einen Ansatz dazu.

- Im Rahmen dieser Arbeit wurde ein Programm erstellt, welches es erlaubt, zeitlich aufgelöste experimentelle Metabolomdaten organismenübergreifend mit theoretischen Netzwerktopologien zusammenzuführen und für statistische Analysen verfügbar zu machen.

Literaturverzeichnis

[Altschul et al. 1990] Altschul, S. F. ; Gish, W. ; Miller, W. ; Myers, E. W. ; Lipman, D. J.: Basic local alignment search tool. In: *Journal of Molecular Biology* 215 (1990), October, Nr. 3, S. 403–410

[Bairoch und Apweiler 2000] Bairoch, Amos ; Apweiler, Rolf: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. In: *Nucleic Acids Research* 28 (2000), Nr. 1, S. 45–48

[Batagelj und Mrvar 1998] Batagelj, V. ; Mrvar, A.: Pajek: Program for large network analysis. In: *Connections* 21 (1998), Nr. 2, S. 47–57

[van den Berg et al. 2006] Berg, R.A. van den ; Hoefsloot, H.C.J. ; Westerhuis, J.A. ; Smilde, A.K. ; Werf, M.J. van der: Centering, scaling, and transformations: improving the biological information content of metabolomics data. In: *BMC Genomics* 7 (2006), Nr. 1, S. 142

[Besemer et al. 2001] Besemer, J. ; Lomsadze, A. ; Borodovsky, M.: GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. In: *Nucleic Acids Research* 29 (2001), Nr. 12, S. 2607–2618

[Bino et al. 2004] Bino, R.J. ; Hall, R.D. ; Fiehn, O. ; Kopka, J. ; Saito, K. ; Draper, J. ; Nikolau, B.J. ; Mendes, P. ; Roessner-Tunali, U. ; Beale, M.H. et al.: Potential of metabolomics as a functional genomics tool. In: *Trends in Plant Science* 9 (2004), Nr. 9, S. 418–425

[Bray 2003] Bray, D.: Molecular Networks: The Top-Down View. In: *Science* 301 (2003), Nr. 5641, S. 1864–1865

[Camacho et al. 2005] Camacho, D. ; Fuente, A. ; Mendes, P.: The origin of correlations in metabolomics data. In: *Metabolomics* 1 (2005), Nr. 1, S. 53–63

[Caspi et al. 2006] Caspi, R. ; Foerster, H. ; Fulcher, C. A. ; Hopkinson, R. ; Ingraham, J. ; Kaipa, P. ; Krummenacker, M. ; Paley, S. ; Pick, J. ; Rhee, S. Y. ; Tissier, C. ; Zhang, P. ; Karp, P. D.: MetaCyc: a multiorganism database of metabolic pathways and enzymes. In: *Nucleic Acids Research* 34 (2006), S. 511–516

[Covert et al. 2001] Covert, M.W. ; Schilling, C.H. ; Famili, I. ; Edwards, J.S. ; Goryanin, I.I. ; Selkov, E. ; Palsson, B.O.: Metabolic modeling of microbial strains *in silico*. In: *Trends in Biochemical Sciences* 26 (2001), Nr. 3, S. 179–186

[Csete und Doyle 2004] Csete, M. ; Doyle, J.: Bow-ties, metabolism and disease. In: *Trends in Biotechnology* 22 (2004), Nr. 9, S. 446–450

[Cypionka 2005] Cypionka, H.: *Grundlagen der Mikrobiologie*. Springer, 2005

[Delcher et al. 1999] Delcher, A. L. ; Harmon, D. ; Kasif, S. ; White, O. ; Salzberg, S. L.: Improved microbial gene identification with GLIMMER. In: *Nucleic Acids Research* 27 (1999), December, Nr. 23, S. 4636–4641

[Delcher et al. 2007] Delcher, A.L. ; Bratke, K.A. ; Powers, E.C. ; Salzberg, S.L.: Identifying bacterial genes and endosymbiont DNA with Glimmer. In: *Bioinformatics* 23 (2007), Nr. 6, S. 673

[Dominguez et al. 1998] Dominguez, H. ; Rollin, C. ; Guyonvarch, A. ; Guerquin-Kern, J.L. ; Cocaign-Bousquet, M. ; Lindley, N.D.: Carbon-flux distribution in the central metabolic pathways of *Corynebacterium glutamicum* during growth on fructose. In: *European Journal of Biochemistry* 254 (1998), Nr. 1, S. 96–102

[Dunn et al. 2005] Dunn, W. B. ; Bailey, N. J. ; Johnson, H. E.: Measuring the metabolome: current analytical technologies. In: *The Analyst* 130 (2005), Nr. 5, S. 606–25

[Edwards et al. 2002] Edwards, J.S. ; Covert, M. ; Palsson, B.: Metabolic modelling of microbes. The flux balance approach. In: *Environmental Microbiology* 4 (2002), Nr. 3, S. 133–140

[Eggeling und Bott 2005] Eggeling, L. ; Bott, M.: *Handbook of Corynebacterium glutamicum.* CRC Press Taylor and Francis Group, 2005

[Fiehn 2001] Fiehn, O.: Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. In: *Comparative and Functional Genomics* 2 (2001), Nr. 3, S. 155–168

[Fiehn et al. 2000] Fiehn, O. ; Kopka, J. ; Dormann, P. ; Altmann, T. ; Trethewey, R. N. ; Willmitzer, L.: Metabolite profiling for plant functional genomics. In: *Nature Biotechnology* 18 (2000), Nr. 11, S. 1157–61

[Frimmersdorf 2005] Frimmersdorf, E.: *Metabolomanalyse von Corynebacterium glutamicum nach Kultivierung auf verschiedenen Kohlenstoff- und Stickstoffquellen (Diplomarbeit).* 2005

[Gerstmeir et al. 2003] Gerstmeir, R. ; Wendisch, V.F. ; Schnicke, S. ; Ruan, H. ; Farwick, M. ; Reinscheid, D. ; Eikmanns, B.J.: Acetate metabolism and its regulation in *Corynebacterium glutamicum*. In: *Journal of Biotechnology* 104 (2003), Nr. 1-3, S. 99–122

[Goodacre et al. 2004] Goodacre, R. ; Vaidyanathan, S. ; Dunn, W. B. ; Harrigan, G. G. ; Kell, D. B.: Metabolomics by numbers: acquiring and understanding global metabolite data. In: *Trends in Biotechnology* 22 (2004), Nr. 5, S. 245–52

[Goto et al. 2002] 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

[Hartmann 2007] Hartmann, K.: *Modellierung stationärer Zustände von metabolischen Netzwerken: Methoden, Anwendungen, Thermodynamik (Dissertation).* 2007

[Hastie et al. 2001] Hastie, T. ; Tibshirani, R. ; Friedman, J.: *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2001

[Hatakeyma et al. 2000] Hatakeyma, K. ; Goto, M. ; Kobayashi, M. ; Terasawa, M. ; Yukawa, H.: Analysis of Oxidation Sensitivity of Maleate cis-trans

Isomerase from *Serratia marcescens*. In: *Bioscience, Biotechnology, and Biochemistry* 64 (2000), Nr. 7, S. 1477–1485

[Hayashi et al. 2002] Hayashi, M. ; Mizogushi, H. ; Shiraishi, N. ; Obayashi, M. ; Nakagawa, S. ; Imai, J. ; Watanabe, S. ; Ota, T. ; Ikeda, M.: Transcriptome Analysis of Acetate Metabolism in *Corynebacterium glutamicum* Using a Newly Developed Metabolic Array. In: *Bioscience, Biotechnology, and Biochemistry* 66 (2002), Nr. 6, S. 1337–1344

[Hermann 2003] Hermann, T.: Industrial production of amino acids by coryneform bacteria. In: *Journal of Biotechnology* 104 (2003), Nr. 1-3, S. 155–172

[Horning und Horning 1971] Horning, E. C. ; Horning, M. G.: Metabolic Profiles: Gas-Phase Methods for Analysis of Metabolites. In: *Clinical Chemistry* 17 (1971), Nr. 8, S. 802–809

[Hulo et al. 2006] Hulo, N. ; Bairoch, A. ; Bulliard, V. ; Cerutti, L. ; De Castro, E. ; Langendijk-Genevaux, P.S. ; Pagni, M. ; Sigrist, C.J.A.: The PROSITE database. In: *Nucleic Acids Research* 34 (2006), S. D227–D230

[Ikeda und Nakagawa 2003] Ikeda, M. ; Nakagawa, S.: The *Corynebacterium glutamicum* genome: features and impacts on biotechnological processes. In: *Applied Microbiology and Biotechnology* 62 (2003), Nr. 2, S. 99–109

[Ishii et al. 2007] Ishii, N. ; Nakahigashi, K. ; Baba, T. ; Robert, M. ; Soga, T. ; Kanai, A. ; Hirasawa, T. ; Naba, M. ; Hirai, K. ; Hoque, A. ; Yee Ho, P. ; Kakazu, Y. ; Sugawara, K. ; Igarashi, S. ; Harada, S. ; Masuda, T. ; Sugiyama, N. ; Togashi, T. ; Hasegawa, M. ; Takai, Y. ; Yugi, K. ; Arakawa, K. ; Iwata, N. ; Toya, Y. ; Nakayama, Y. ; Nishioka, T. ; Shimizu, K. ; Mori, H. ; Tomita, M.: Multiple High-Throughput Analyses Monitor the Response of *E. coli* to Pertubations. In: *Science* 316 (2007), Nr. 593

[Jeong et al. 2000] Jeong, H. ; Tombor, B. ; Albert, R. ; Oltvai, Z.N. ; Barabasi, A. L.: The large-scale organization of metabolic models. In: *Nature* 407 (2000), Nr. 6804, S. 651–4

[Jungnickel 2002] Jungnickel, D.: *Graphs, Network and Algorithm.* Springer-Verlag, Berlin, 2002

[Kalinowski et al. 2003] Kalinowski, J. ; Bathe, B. ; Bartels, D. ; Bischoff, N. ; Bott, M. ; Burkovski, A. ; Dusch, N. ; Eggeling, L. ; Eikmanns, B. J. ; Gaigalat, L. ; Goesmann, A. ; Hartmann, M. ; Huthmacher, K. ; Krämer, R. ; Linke, B. ; McHardy, A. C. ; Meyer, F. ; Mockel, B. ; Pfefferle, W. ; Puhler, A. ; Rey, D. A. ; Ruckert, C. ; Rupp, O. ; Sahm, H. ; Wendisch, V. F. ; Wiegrabe, I. ; Tauch, A.: The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. In: *Journal of Biotechnology* 104 (2003), Nr. 1-3, S. 5–25

[Kanehisa et al. 2004] Kanehisa, M. ; Goto, S. ; Kawashima, S. ; Okuno, Y. ; Hattori, M.: The KEGG resource for deciphering the genome. In: *Nucleic Acids Research* 32 (2004), Nr. 90001, S. 277–280

[Kell 2004] Kell, D. B.: Metabolomics and systems biology: making sense of the soup. In: *Current Opinion in Microbiology* 7 (2004), Nr. 3, S. 296–307

[Kitano 2002] Kitano, H.: Systems Biology: A Brief Overview. In: *Science* 295 (2002), Nr. 5560, S. 1662–1664

[Kitano 2004] Kitano, H.: Biological robustness. In: *Nature Reviews Genetics* 5 (2004), Nr. 11, S. 826–837

[Koek et al. 2006] Koek, M.M. ; Muilwijk, B. ; Werf, M.J. van der ; Hankemeier, T.: Microbial metabolomics with gas chromatography/mass spectrometry. In: *Analytical Chemistry* 78 (2006), Nr. 4, S. 1272–1281

[Kose et al. 2001] Kose, F. ; Weckwerth, W. ; Linke, T. ; Fiehn, O.: Visualizing plant metabolomic correlation networks using clique-metabolite matrices. In: *Bioinformatics* 17 (2001), Nr. 12, S. 1198–208. – 1367-4803 (Print) Journal Article

[Krebs und Johnson 1937] Krebs, H.A. ; Johnson, W.A.: The role of citric acid in intermediate metabolism in animal tissues. In: *Enzymologica* 4 (1937), S. 148–156

[Krieger et al. 2004] Krieger, C.J. ; Zhang, P. ; Mueller, L.A. ; Wang, A. ; Paley, S. ; Arnaud, M. ; Pick, J. ; Rhee, S.Y. ; Karp, P.D.: MetaCyc: a

multiorganism database of metabolic pathways and enzymes. In: *Nucleic Acids Research* (2004)

[Kruse et al. 1993] Kruse, F.A. ; Lefkoff, A.B. ; Boardman, J.W. ; Heidebrecht, K.B. ; Shapiro, A.T. ; Barloon, P.J. ; Goetz, A.F.H.: The Spectral Image Processing System (SIPS)- Interactive visualization and analysis of imaging spectrometer data. In: *Remote Sensing of Environment* 44 (1993), Nr. 2, S. 145–163

[Le Novère et al. 2005] Le Novère, N. ; Finney, A. ; Hucka, M. ; Bhalla, U.S. ; Campagne, F. ; Collado-Vides, J. ; Crampin, E.J. ; Halstead, M. ; Klipp, E. ; Mendes, P. et al.: Minimum information requested in the annotation of biochemical models (MIRIAM). In: *Nature Biotechnology* 23 (2005), S. 1509–1515

[Liebl 1991] Liebl, W.: The genus Corynebacterium - nonmedical. In: *The Procaryotes* 2 (1991), S. 1157–1171

[Ma und Zeng 2003a] Ma, H. ; 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

[Ma und Zeng 2003b] 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

[Massart et al. 1997] Massart, D.L. ; B.G.M, Vandeginste ; L.M.C., Buydens: *Handbook of chemometrics and Qualimetrics, Part A*. Elsevier, Amsterdam, 1997

[Mavrovouniotis 1991] Mavrovouniotis, M.L.: Estimation of standard Gibbs energy changes of biotransformations. In: *Journal of Biological Chemistry* 266 (1991), Nr. 22, S. 14440–14445

[Mori und Shiio 1987] Mori, M. ; Shiio, I.: Phosphoenolpyruvate: sugar transferase systems and sugar metabolism in *Brevibacterium flavum*. In: *Agricultural and biological chemistry* 51 (1987), S. 2671–2678

[Muffler et al. 2002] Muffler, A. ; Bettermann, S. ; Haushalter, M. ; Hörlein, A. ; Neveling, U. ; Schramm, M. ; Sorgenfrei, O.: Genome-wide transcription profiling of *Corynebacterium glutamicum* after heat shock and during growth on acetate and glucose. In: *Journal of Biotechnology* 98 (2002), Nr. 2-3, S. 255–268

[Palsson et al. 2003] Palsson, B.O. ; Price, N.D. ; Papin, J.A.: Development of network-based pathway definitions: the need to analyze real metabolic networks. In: *Trends in Biotechnology* 21 (2003), Nr. 5, S. 195–198

[Pearl 1984] Pearl, J.: *Heuristics: intelligent search strategies for computer problem solving*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1984

[Rahman und Schomburg 2006] Rahman, S. A. ; Schomburg, D.: Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. In: *Bioinformatics* (2006). – 1367-4803 (Print) Journal article

[Rahman et al. 2005] Rahman, S.A. ; Advani, P. ; Schunk, R. ; Schrader, R. ; Schomburg, D.: Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). In: *Bioinformatics* 21 (2005), Nr. 7, S. 1189–1193

[Ravasz et al. 2002] Ravasz, E. ; Somera, AL ; Mongru, D.A. ; Oltvai, Z.N. ; Barabasi, A.L.: Hierarchical Organization of Modularity in Metabolic Networks. In: *Science* 297 (2002), Nr. 5586, S. 1551–1555

[Roessner et al. 2001] Roessner, U. ; Luedemann, A. ; Brust, D. ; Fiehn, O. ; Linke, T. ; Willmitzer, L. ; Fernie, A.R.: Metabolic Profiling Allows Comprehensive Phenotyping of Genetically or Environmentally Modified Plant Systems. In: *The Plant Cell Online* 13 (2001), S. 11–29

[Roessner et al. 2000] Roessner, U. ; Wagner, C. ; Kopka, J. ; Trethewey, R.N. ; Willmitzer, L.: Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. In: *The Plant Journal* 23 (2000), Nr. 1, S. 131–142

[Sauer et al. 2007] Sauer, U. ; Heinemann, M. ; Zamboni, N.: Getting Closer to the Whole Picture. In: *Science* 316 (2007), Nr. 5824, S. 550

[Sauter et al. 1991] Sauter, H. ; Lauer, M. ; Fritsch, H.: Metabolic profiling of plants - a new diagnostic technique. In: *Synthesis and Chemistry of Agrochemicals II* (1991), S. 288–299

[Schilling et al. 2002] Schilling, C.H. ; Covert, M.W. ; Famili, I. ; Church, G.M. ; Edwards, J.S. ; Palsson, B.O.: Genome-Scale Metabolic Model of Helicobacter pylori 26695. In: *Journal of Bacteriology* 184 (2002), Nr. 16, S. 4582–4593

[Schilling et al. 2001] Schilling, C.H. ; Edwards, J.S. ; Palsson, B.O.: Toward Metabolic Phenomics: Analysis of Genomic Data Using Flux Balances. In: *Introduction to Bioengineering* 15 (2001), Nr. 3, S. 288–295

[Schilling und Palsson 1998] Schilling, C.H. ; Palsson, B.O.: The underlying pathway structure of biochemical reaction networks. In: *Proceedings of the National Academy of Sciences* 95 (1998), Nr. 8, S. 4193–4198

[Schomburg et al. 2002] Schomburg, I. ; Chang, A. ; Schomburg, D: BRENDA, enzyme data and metabolic information. In: *Nucleic Acids Research* 30 (2002), Nr. 1, S. 47–49

[Schweingruber 1983] Schweingruber, F.H.: *Der Jahrring. Standort, Methodik, Zeit und Klima in der Dendrochronologie.* Verlag Paul Haupt, 1983

[Stein 1999] Stein, S.E.: An integrated method for spectrum extraction and compound identification from GC/MS data. In: *Journal of the American Society for Mass Spectrometry* 10 (1999), S. 770–781

[Stelling et al. 2002] Stelling, J. ; Klamt, S. ; Bettenbrock, K. ; Schuster, S. ; Gilles, E. D.: Metabolic network structure determines key aspects of functionality and regulation. In: *Nature* 420 (2002), Nr. 6912, S. 190–193

[Steuer et al. 2002] Steuer, R. ; Kurths, J. ; Daub, C.O. ; Weise, J. ; Selbig, J.: The mutual information: detecting and evaluating dependencies between variables. In: *Bioinformatics* 18 (2002), Nr. 2, S. 231–240

[Steuer et al. 2003] Steuer, R. ; Kurths, J. ; Fiehn, O. ; Weckwerth, W.: Observing and interpreting correlations in metabolomic networks. In: *Bioinformatics* 19 (2003), Nr. 8, S. 1019–26

[Strelkov et al. 2004] Strelkov, S. ; Elstermann, M. von ; Schomburg, D.: Comprehensive analysis of metabolites in *Corynebacterium glutamicum* by gas chromatography/mass spectrometry. In: *Journal of Biological Chemistry* 385 (2004), Nr. 9, S. 853–861

[Tweeddale et al. 1998] Tweeddale, H. ; Notley-McRobb, L. ; Ferenci, T.: Effect of Slow Growth on Metabolism of *Escherichia coli*, as Revealed by Global Metabolite Pool ("Metabolome") Analysis. In: *Journal of Bacteriology* 180 (1998), Nr. 19, S. 5109–5116

[Urbanczyk-Wochniak et al. 2003] Urbanczyk-Wochniak, E. ; Luedemann, A. ; Kopka, J. ; Selbig, J. ; Roessner-Tunali, U. ; Willmitzer, L. ; Fernie, A.R.: Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. In: *EMBO Reports* 4 (2003), Nr. 10, S. 989–993

[Ward 1963] Ward, J.H.: Hierarchical grouping to optimize an objective function. In: *Journal of the American Statistical Association* 58 (1963), Nr. 301, S. 236–244

[Weckwerth 2003] Weckwerth, W.: Metabolomics in Systems Biology. In: *Annual Review of Plant Biology* 54 (2003), Nr. 1, S. 669–689

[Weckwerth et al. 2004] Weckwerth, W. ; Loureiro, M.E. ; Wenzel, K. ; Fiehn, O.: Differential metabolic networks unravel the effects of silent plant phenotypes. In: *Proceedings of the National Academy of Sciences* 101 (2004), Nr. 20, S. 7809–7814

[Wendisch et al. 2000] Wendisch, V.F. ; Graaf, A.A. de ; Sahm, H. ; Eikmanns, B.J.: Quantitative Determination of Metabolic Fluxes during Coutilization of Two Carbon Sources: Comparative Analyses with *Corynebacterium glutamicum* during Growth on Acetate and/or Glucose. In: *Journal of Bacteriology* 182 (2000), Nr. 11, S. 3088

[Yamazaki et al. 2003] Yamazaki, M. ; Nakajima, J. ; Yamanashi, M. ; Sugiyama, M. ; Makita, Y. ; Springob, K. ; Awazuhara, M. ; Saito, K.: Metabolomics and differential gene expression in anthocyanin chemo-varietal forms of *Perilla frutescens*. In: *Phytochemistry* 62 (2003), Nr. 6, S. 987–995