Kodierung enzymatischer Reaktionen

Inaugural - Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Markus Leber

aus Winterberg

Köln 2008

Berichterstatter:	Prof. Dr. D. Schomburg
	Prof. Dr. R. Schrader
Tag der mündlichen Prüfung:	25.04.2008

Danksagung

Ich möchte mich bei allen bedanken, die zum Gelingen dieser Arbeit beigetragen haben.

Mein besonderer Dank gilt Prof. Dr. Dietmar Schomburg für das interessante Thema, seine ständige Bereitschaft zum fachlichen Dialog und die konstruktive Unterstützung in den letzten Jahren. Ebenso für den gewährten Freiraum, der die Entwicklung neuer Verfahren ermöglichte.

Prof. Dr. Rainer Schrader möchte ich für die freundliche Übernahme des Zweitgutachtens danken und für die Hilfsbereitschaft während meines CUBIC Studienganges und in den darauf folgenden Jahren.

Prof. Dr. Ulrich Deiters und PD Dr. Karsten Niefind danke ich für die Übernahme des Vorsitzes bzw. Beisitzes bei meiner Disputation.

Besonders hervorheben möchte ich die Unterstützung durch Dr. Volker Egelhofer und Dr. Ida Schomburg, die durch die Konstruktion und Korrektur tausender Reaktionen entscheidend zum Erfolg dieser Arbeit beigetragen haben. Ich möchte mich für diese großartige Leistung bedanken.

Vielen Dank ebenso an meine Bürokollegen Dr. Silke Schrader und Dr. Kai Hartmann, die mir stets unterstützend zur Seite standen und mir in besonderer Weise Innovation und Ausdauer gegeben haben. Dr. Silke Schrader danke ich darüber hinaus für das Korrekturlesen meiner Arbeit. Herzlichen Dank auch an Michael Zimmermann für die systemadministrative Betreuung und die aufmunternden Kommentare. Darüber hinaus danke ich allen derzeitigen und ehemaligen Mitarbeitern des Arbeitskreises, die hier nicht alle aufgeführt werden können, für die stete Hilfsbereitschaft, die angenehme Arbeitsatmosphäre und das kollegiale Miteinander.

Inhaltsverzeichnis

Summary

Classification of enzymes is performed by committees of the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Biochemistry and Molecular Biology (IUBMB). The assignment of EC-numbers is a complex procedure, which can not be performed by single individuals. False assignments and enzymes, which were grouped for historical reasons or physiological aspects, lead to steadily updates of the enzyme list. New enzymes are added and false classified enzymes are transferred or deleted completely. This extensive procedure was the motivation for the present work to investigate an automated system for the characterization of enzymatic reactions. The Dugundji-Ugi-model describes reactions by mathematical operators named transition or reaction matrices (R-matrices). R-matrices represent electron shift patterns occurring during reactions and allow an objective view on the reaction core. Nevertheless the model has not been used excessively within the last years, because it is difficult to handle without an assignment of the educt atoms to the product atoms. For this reason R-matrices have been generated manually in the past.

In the present work a program has been developed, which calculates R-matrices automatically only on the basis of the educt and product molecules. The atom mapping problem is tackled by a MCS algorithm (Maximal Common Subgraph), which determines maximal common substructures of two input molecules. This newly developed algorithm is a flexible approach, which is able to reproduce complex changes of the structure and allows the comparison of large metabolites as well. The MCS algorithm is used to compare each educt molecule to each product molecule. In the next step the maximal common substructures of the different comparisons are combined and assessed by a ranking system, which calculates all possible atom-atom mappings. Due to the complexity of various biochemical reactions, it was necessary to extend the system. Procedures were integrated to find local symmetries, aromatic rings and cofactors. Small molecules and remaining structures are considered by further MCS methods. This system helps to avoid false assignments and to improve the run time.

On the basis of a complete atom mapping it is comparatively simple to calculate the reaction matrix. More complex is the identification of identical R-matrices. The comparison of R-matrices is circumstantial according to the n! possible permutations. Due to the complex structure of biochemical electron shift patterns it was necessary to develop a new canonization algorithm, which constructs a representative R-matrix for each reaction. Finally the canonical matrices are converted to R-strings. R-strings can be easily compared and allow a new grouping of enzymatic reactions according to their electron shift patterns.

The R-matrix calculator is tested for a set of 3209 enzymatic reactions, which covers 228 of 229 subsubclasses. A main objective of the analysis was the comparison of the Dugundji-Ugi-Model and the EC-classification system. The subsubclasses of the EC-classification system were used as a reference point. First the R-matrix identity within the subsubclasses was examined. In the most cases the R-matrices within a subsubclass are homogeneous. Nevertheless the system also reveals subsubclasses with inconsistent electron shift patterns.

In another examination R-matrices of different subsubclasses were compared and grouped. For this purpose for each subsubclass the R-matrix with the highest occurrence was investigated and used for R-matrix comparisons. 121 subsubclasses have an unique R-matrix. This means enzymes of these subsubclasses can be recognized by the corresponding R-matrix with high probability. On the other hand 107 subsubclasses do not have an unique R-matrix. They form 26 groups, whereas each group share an identical R-matrix. The R-matrices of the largest groups show basic electron shift patterns. Although these subsubclasses have identical R-matrices, they sometimes belong to different main classes within the EC-classification system. Obviously the electron shift patterns differ in many aspects from the EC-classification system.

The results demonstrate the properties of the Dugundji-Ugi-model as a rational system, which is not constricted by prejudices of our experience and allows an objective view on biochemical reactions. It could be used to support the EC-classification system, especially as it also considers the chemical properties of the overall reaction. On the basis of this model subsubclasses can be identified, which can be partitioned into smaller groups or can be joined to new groups of identical reaction cores.

Zusammenfassung

Die Einteilung enzymatischer Reaktionen erfolgt auf Basis des EC-Klassifikationssystems. Die Einordnung neuer Enzyme ist hochkomplex und erfordert die Absprache von verschiedenen Enzymkommissionen der Organisationen IUPAC und IUBMB. Dieses zeitaufwendige Verfahren war die Motivation für die vorliegende Arbeit, ein automatisiertes System für die Charakterisierung enzymatischer Reaktionen zu entwickeln, das auf dem Dugundji-Ugi-Modell basiert. Hierbei werden Reaktionen durch mathematische Operatoren beschrieben, die als Reaktionsmatrizen (R-Matrizen) bezeichnet werden und das Elektronentransfermuster einer Reaktion kodieren. R-Matrizen enthalten die Information, welche Bindungen gespalten werden oder entstehen und welche Atome an der Reaktion beteiligt sind. Das Errechnen der R-Matrizen erfordert allerdings eine Atomzuordnung der Eduktatome auf die Produktatome. Diese Zuordnung wurde in der Vergangenheit immer mit hohem manuellem Aufwand erstellt.

Im Rahmen dieser Arbeit wurde ein Programm entwickelt, das die R-Matrizen nur anhand der Edukt- und Produktmoleküle errechnet und ohne manuelle Unterstützung auskommt. Kern des Verfahrens ist ein MCS-Algorithmus (Maximal Common Subgraph), der die maximalen gemeinsamen Substrukturen von 2 Molekülen errechnet. Dieser neu entwickelte Algorithmus ist sehr flexibel und kann daher auch komplexe Veränderungen in der Molekülstruktur nachvollziehen. Er wird dazu verwendet, die Eduktmoleküle mit den Produktmolekülen zu vergleichen. Die maximalen gemeinsamen Substrukturen aus den Molekülvergleichen werden kombiniert und durch ein Rankingsystem bewertet, das alle möglichen Atomzuordnungen errechnet. Bedingt durch die Komplexität vieler biochemischer Reaktionen, war es erforderlich, das Verfahren um weitere Algorithmen zu erweitern, um falsche Zuordnungen zu vermeiden und die Laufzeit zu verbessern. So wurden Verfahren zur Erkennung von lokalen Symmetrien, aromatischen Ringsystemen und Kofaktoren integriert und eine weitere Molekülvergleichsebene eingeführt. Mit Hilfe dieses Systems von verschiedenen Algorithmen wird eine vollständige Atomzuordnung generiert.

Auf Grundlage einer vollständigen Atomzuordnung ist die Berechnung der R-Matrizen schließlich sehr einfach. Ein weiteres Problem stellte allerdings der Vergleich der R-Matrizen dar, weil von jeder R-Matrix n! mögliche Permutationen erzeugt werden können. Der Vergleich erfolgt mit Hilfe eines neuen Kanonisierungsalgorithmus, der aus der Vielzahl von Permutationen eine repräsentative R-Matrix selektiert. Die Überführung der R-Matrizen in R-Strings vereinfacht schließlich den Vergleich und die Gruppierung von enzymatischen Reaktionen anhand ihres Elektronentransfermusters.

Auf der Grundlage eines Datensets, das 228 der 229 definierten Subsubklassen abdeckt, wurden die R-Matrizen von 3209 enzymatischen Reaktionen automatisch errechnet. Neben der Analyse der R-Matrizen selbst, standen bei der Auswertung die Beziehungen zwischen dem Dugundji-Ugi-Modell und dem EC-Klassifikationssystem im Vordergrund. Als Bezugspunkt wurden die Subsubklassen des EC-Klassifikationssystems gewählt, die bereits soweit spezifiziert sind, dass hier eine hohe Übereinstimmung beider Systeme zu erwarten war. So wurden die Subsubklassen auf ihre Homogenität in ihren Elektronentransfermustern untersucht. Die häufigsten R-Matrizen jeder Subsubklasse wurden verglichen und Subsubklassen mit identischen R-Matrizen gruppiert.

Die Elektronentransfermuster innerhalb der Subsubklassen erwiesen sich in der Mehrzahl als homogen. Allerdings wurden auch Subsubklassen mit geringer Übereinstimmung in den Elektronentransfermustern gefunden, für die eine höhere Homogenität erwartet worden wäre. Der R-Matrixvergleich der verschiedenen Subsubklassen ergab, dass 121 Subsubklassen bereits ein spezifisches Elektronentransfermuster besitzen. 107 Subsubklassen bilden hingegen Gruppen verschiedener Größe. Einige Gruppen bestehen aus Subsubklassen von verschiedenen Hauptklassen, obwohl sie nach ihren Elektronentransfermustern einen identischen Reaktionskern besitzen.

Die Ergebnisse machen die Eigenschaften des Dugundji-Ugi-Modells deutlich, das sehr rational und unvoreingenommen die wesentlichsten Eigenschaften einer Reaktion beschreibt. Ähnlich, wie das EC-Klassifikationssystem, betrachtet es aber die chemischen Eigenschaften der Gesamtreaktion. Es könnte daher die Enzymklassifikation unterstützen, indem es aufzeigt, welche Subsubklassen feiner untergliedert oder zu größeren Gruppen zusammengefasst werden könnten.

Abkürzungsverzeichnis

Enzymklassifikation:

EC	Enzyme Commission
IUPAC	International Union of Pure and Applied Chemistry
IUBMB	International Union of Biochemistry and Molecular Biology
NC-IUBMB	Nomenclature Committee of the International Union of Biochemistry and Molecular Biology
JCBN	Joint Commission on Biochemical Nomenclature

Datenbanken:

BRENDA	Braunschweiger Enzymdatenbank
KEGG	Kyoto Encyclopedia of Genes and Genomes

Begriffe aus der Graphenthreorie und dem Dugundji-Ugi-Modell:

MCS	Maximum Common Subgraph / Substructure
c-MCS	*connected*-Maximum Common Subgraph / Substructure
BE-Matrix	Bindungs-Elektronen Matrix
R-Matrix	Reaktionsmatrix

Metabolite:

H_2O	Wasser
NH_3	Ammoniak
H_2	elementarer Wasserstoff
H^+	Proton
CoA	Coenzym A
NDP	Nukleosid-Diphosphat
NTP	Nukleosid-Triphosphat
AMP	Adenosin-Monophosphat
ADP	Adenosin-Diphosphat
ATP	Adenosint-Triphosphat
FAD	Flavin-Adenin-Dinukleotid
$FADH_2$	Flavin-Adenin-Dinukleotid (reduziert)
FMN	Flavin-Mononukleotid
$FMNH_2$	Flavin-Mononukleotid (reduziert)
NAD	Nikotinamid-Adenin-Dinukleotid
NADH	Nikotinamid-Adenin-Dinukleotid (reduziert)
NADP	Nikotinamid-Adenin-Dinukleotid-Phosphat
NADPH	Nikotinamidadenindinukleotidphosphat (reduziert

1. Einleitung

1.1 Enzyme und ihre Eigenschaften

Enzyme sind Proteine, die biochemische Reaktionen katalysieren. Sie beschleunigen Reaktionen, indem sie die Aktivierungsenergie für eine bestimmte chemische Veränderung selektiv vermindern. Die Geschwindigkeit der Reaktionen wird dabei mindestens um den Faktor eine Million erhöht. Daher würden ohne Enzyme selbst die einfachsten biochemischen Reaktionen ohne wahrnehmbaren Effekt ablaufen (Voegt, 2004).

Enzyme bewirken und kontrollieren die verschiedensten Abläufe und sind der Schlüssel zum Verständnis der Stoffwechselvorgänge in Organismen. Beispielsweise katalysieren Enzyme die Synthese und den Abbau von Biomolekülen oder bewirken die Gewinnung und Umwandlung von Energie (Photosynthese, Glycolyse, Citrat-Zyklus, Atmungskette).

Sie sind aber nicht nur Katalysatoren, sondern fungieren auch als raffinierte Umwandler von Bewegung, als Signal-Verarbeiter, oder als Aktivatoren bzw. Deaktivatoren und nehmen so Einfluss auf nahezu alle dynamischen Prozesse der Zelle. Zum Beispiel sind hunderte von Proteinkinasen in eurkaryontischen Zellen in komplexen Netzwerken der Signalübertragung organisiert (Rodbell, 1980, Li *et al.*, 2006). Diese Proteinkinasen vermitteln und verstärken Signale aus der Zellumgebung in die Zelle. Hierdurch beeinflussen sie den Zellzyklus (Morgan, 2007) und unterstützen die Koordinierung der Zellaktivität. Diese vielfältigen Funktionen von Enzymen machen deutlich, warum Enzyme häufig einer komplexen Regulation unterliegen.

Die Wirkung der Enzyme lässt sich allgemein mit dem Kreisprozess beschreiben. Danach bildet das Enzym mit dem Substrat zunächst einen Enzym-Substrat-Komplex, der durch die Reaktion in einen Enzym-Produkt-Komplex übergeht. Durch Zerfall des Enzym-Produkt-Komplexes wird das Produkt freigesetzt und das Enzym ist für weitere Umsetzungen verfügbar (Koshland, 1958).

Im Detail unterscheiden sich die Mechanismen der Enzyme sehr. Dennoch lassen sich bei starker Vereinfachung die enzymatischen Reaktionen auf wenige katalytische Grundstrategien reduzieren. Eine Strategie beruht auf der *Katalyse durch Annäherung*. Dabei werden die Substrate so an das Enzym gebunden, dass die reaktiven Gruppen der Moleküle eine günstige Orientierung zueinander einnehmen. Eine günstige Orientierung und Konformation der Moleküle kann die Reaktionsgeschwindigkeit deutlich erhöhen (Eisenmesser *et al.*, 2002).

Bei der *kovalenten Katalyse* besitzt das aktive Zentrum des Enzyms eine starke meist nukleophile Gruppe, die temporär kovalent an das Substratmolekül bindet. Andere Enzyme besitzen Metallionen, die als elektrophile Katalysatoren wirken oder nukleophile Gruppen erzeugen. Reaktionen dieser Form werden unter der *Metallionenkatalyse* zusammengefasst. Eine bedeutende Rolle spielt auch die

Säure-Base-Katalyse, wo ein Aminosäurerest des Enzyms die Funktion eines Protonendonors oder -akzeptors übernimmt.

Diese Strategien erlauben die Verminderung der Aktivierungsenergie. Unter der Aktivierungsenergie versteht man die Zusatzenergie, die ein Molekül über den Grundzustand hinaus besitzen muss, um in eine bestimmte chemische Reaktion einzutreten. Die Gibbs-Energie für die Reaktion ändert sich durch den Einsatz der Enzyme jedoch nicht. Vielmehr werden Zwischenreaktionen eingegangen, die eine niedrigere Aktivierungsenergie erfordern. Hierdurch wird das Aufbringen der hohen Aktivierungsenergie umgangen (Fersht, 1985, Garcia-Viloca *et al*., 2004)

Enzyme können sich sehr in ihrem Aufbau unterscheiden. Besteht ein Enzym aus einer Polypeptidkette wird es als Monomer bezeichnet. Monomere können eine oder mehrere Reaktionen katalysieren. Sind Enzyme aus mehreren Polypeptidketten aufgebaut, ist das Enzym ein *Oligomer* und die Polypeptidketten werden als Untereinheiten bezeichnet. Manchmal treten Enzyme als *Multienzymkomplex* auf. Dabei sind die Enzyme miteinander assoziiert und katalysieren eine Kette von aufeinander folgenden Reaktionen (Srivastava und Bernhard, 1986).

Einige Enzyme, die *Holoenzyme*, benötigen Kofaktoren für ihre Funktion. Kofaktoren sind keine Proteine sondern zumeist kleinere organische Moleküle, die mit dem eigentlichen Protein (Apoenzym) assoziiert sind. Ist der Kofaktor kovalent an das Protein gebunden wird er als prosthetische Gruppe bezeichnet. Wird der Kofaktor wie das Substrat in der Reaktion modifiziert, werden die Kofaktoren auch als Kosubstrate bezeichnet. Hierzu gehören Energieträger wie Adenosintriphosphat (ATP) oder Oxidations- bzw. Reduktionsmittel wie beispielsweise Nicotinamidadenindinukleotid (NAD+/NADH).

Die Bedingungen der Umgebung sind häufig ebenfalls entscheidend für die enzymatische Aktivität. Viele Enzyme sind bei einer bestimmten Temperatur und einem bestimmten pH-Wert besonders aktiv. Zum Beispiel zeigen lysosomale Enzyme bei einem pH-Wert von 4 ein Optimum in ihrer Aktivität, während sie im Zytosol teilweise fast inaktiv sind.

Da die enzymatische Aktivität kennzeichnend für Enzyme ist, erfolgte ihre Einteilung anhand ihrer beobachtbaren Wirkungs- und Substratspezifität.

1.2 Klassifikation und Nomenklatur von Enzymen nach IUPAC und IUBMB

Bis zur Mitte des letzten Jahrhunderts gab es keine einheitliche Einteilung oder Benennung von Enzymen. Neu entdeckte Enzyme erhielten ihre Namen durch die jeweiligen Wissenschaftler, die an der Endeckung beteiligt waren. Hierdurch konnte ein Enzym verschiedene Namen erhalten, wenn mehrere Gruppen die Entdeckung für sich in Anspruch nahmen. Umgekehrt erhielten Enzyme

verschiedenen Typs einen ähnlichen Namen. Häufig erhielten Enzyme Namen wie Diaphorase, Zwischenferment oder Katalase. Diese Namen gaben aber keinen Hinweis auf die Natur oder die Funktion des Enzyms.

Im Jahre 1955 beschloss daher die *International Union of Biochemistry* (IUB) eine internationale Enzymkommission ins Leben zu rufen, die ein einheitliches Benennungs- und Ordnungssystem für Enzyme erarbeiten sollte (Historical Introduction, 1992). Diese Kommission wurde 1956 unter dem Namen *International Commission on Enzymes* (Internationale Enzymkommission) gegründet. In enger Zusammenarbeit mit der Nomenklaturkommission der IUPAC (*International Union of Pure and Applied Chemistry*) wurden in den folgenden Jahren systematische Regeln für ein einheitliches Ordnungssystem entwickelt. 1961 erfolgte eine erste Veröffentlichung der neuen Enzymliste mit 712 Einträgen.

In den folgenden Jahren nahm die Anzahl der neu entdeckten Enzyme rasch zu, sodass es immer wieder zu Umstrukturierungen des Enzymkomitees kam. 1977 wurde das *Nomenclature Committee of IUB* (NC-IUB; heute NC-IUBMB) gegründet (Keith *et al.*, 2000). Dieses Komitee arbeitet eng zusammen mit der *IUPAC-IUBMB Joint Commission on Biochemical Nomenclature* (JCBN).

Die Enzymliste unterliegt einer ständigen Überarbeitung. Neue Enzyme werden der Liste hinzugefügt, während andere Einträge gelöscht oder korrigiert werden. Heute sind etwa 4000 Enzyme bekannt. Insgesamt wird geschätzt, dass es um die 10000 Enzyme geben könnte.

1.2.1 Klassifikation und Nomenklatur von Enzymen anhand der katalysierten Reaktion

Für die Nomenklatur und Klassifikation wurden von der Enzymkommission gemeinsame Richtlinien erarbeitet, die im Wesentlichen 3 Prinzipien verfolgen (Enzyme Nomenclature, 1992).

Nach dem ersten Prinzip erhalten einzelne Enzyme in ihrem systematischen Enzymnamen die Endung „-ase“. Wird eine Reaktion dagegen von mehreren Enzymen katalysiert, die in einem System organisiert sind, wird die Bezeichnung „System“ oder „Komplex“ in den Namen aufgenommen. Ein Beispiel hierfür ist der Pyruvat-Dehydrogenase-Komplex, der Pyruvat decarboxyliert und Acetyl-CoA erzeugt. In diesem System sind die Enzyme Pyruvat-Dehydrogenase, Dihydrolipoamid-Acetyltransferase und Lipoamiddehydrogenase organisiert. Andere Beispiele für Enzymsysteme sind das Succinat-Oxidase-System, 2-Oxoglutarat-Dehydrogenase-System und das Fettsäuresynthasesystem.

Das zweite Prinzip betrachtet die katalysierte Reaktion als eigentlichen Kern der Enzymbenennung und Klassifizierung. Die katalysierte chemische Reaktion ist das spezifische Merkmal von Enzymen und bietet sich daher zur Unterscheidung und Klassifikation an. Auf Basis der formalen

Reaktionsgleichung werden die Veränderungen der Produktmoleküle gegenüber den Eduktmolekülen abgeleitet. Über welche Übergangszustände und Zwischenprodukte der Reaktionsmechanismus verläuft wird nicht berücksichtigt. Somit lässt sich das Klassifikationssystem der Enzyme ebenso als ein Ordnungssystem chemischer Reaktionen auffassen. Dieses Verfahren setzt voraus, dass die enzymatische Reaktion vollständig beschrieben ist. Enzyme, für die nur ein Teil ihrer Reaktion bekannt ist, können zunächst nicht benannt und klassifiziert werden, bis eine Gesamtreaktion formuliert werden kann. Neben der chemischen Reaktion wurden weitere Möglichkeiten der Klassifikation in Betracht gezogen, wie die chemischen Eigenschaften des Enzyms. Beispielsweise könnte man einen Teil der Enzyme anhand ihrer aktiven Zentren oder prosthetischen Gruppen einteilen. Allerdings sind nicht alle aktiven Zentren aufgeklärt und nur ein kleiner Teil der Enzyme besitzt prosthetische Gruppen. Dennoch werden diese Kriterien in Ausnahmefällen zur Klassifikation hinzugezogen, um größere Enzymgruppen mit identischer Reaktion zu unterscheiden.

Nach dem Schema des dritten Prinzips werden größere Gruppen von Enzymen anhand des Typs der Reaktion eingeordnet. Der Reaktionstyp dient zusammen mit den Namen der beteiligten Substrate als Grundlage für die Enzymbenennung. Auf diese Weise erhält der Name der einzelnen Enzyme eine Beschreibung und Klassifikation der Reaktion. Beispielsweise lässt sich am Namen der Homoserine-Dehydrogenase ableiten, dass dieses Enzym Homoserin dehydriert. Aus dem Namen Lysine-Decarboxylase wird hingegen ersichtlich, dass die Carboxylgruppe des Lysin abgespalten wird und dieses Enzym zur Gruppe der Decarboxylasen gehört.

In manchen Fällen war es erforderlich, spezielle Regeln zu entwickeln. So können einzelne Enzyme auch sehr komplexe Reaktionen katalysieren, die über mehrere Teilschritte unterschiedlichen Reaktionstyps verlaufen. Dabei können einige Zwischenschritte spontan ohne Beteiligung des Enzyms ablaufen, während andere die Unterstützung des Enzyms als Katalysator benötigen. In diesem Fall wird nur die erste Teilreaktion für die Klassifikation herangezogen, die für die nachfolgenden Reaktionen von essentieller Bedeutung ist.

Wichtig für die Klassifikation und Namensgebung ist auch die Frage nach der Richtung der Reaktionen. Zur Vereinfachung wurde die Reaktionsrichtung für eine Enzymklasse vereinheitlicht, auch wenn diese Richtung nicht für alle Enzyme nachgewiesen werden konnte.

1.2.2 Systematische Namen und Trivialnamen

Nach dem Nomenklatursystem der ersten Enzymkommission erhalten Enzyme einen systematischen Namen und einen Trivialnamen.

Bei den Trivialnamen handelt es sich häufig um ältere Namen, die immer noch weit verbreitet sind. Aus historischen Gründen wird in einigen Fällen weiterhin der Trivialname verwendet und auf systematische Namen verzichtet, wie bei Trypsin oder Chymotrypsin. Obwohl Trivialnamen in einigen Fällen nicht erkennen lassen, dass es sich bei dieser Substanz um ein Enzym handelt, sind sie in der Regel kürzer und einfacher anzuwenden als die systematischen Namen. Aus diesem Grund werden sie auch von der Enzymkommission weiter verwendet und folgen in der Beschreibung unmittelbar auf die Klassifikationsnummern.

Die systematischen Namen werden nach definierten Regeln aus dem Reaktionstyp und den Namen der beteiligten Substrate gebildet. Beispielsweise wurde für das Enzym mit dem Trivialnamen *Alkohol-Dehydrogenase* der systematische Name *Alkohol:NAD+ Oxidoreductase* gewählt, da es ein Alkoholmolekül unter Reduktion von NAD+ zu einem Aldehydmolekül umwandeln kann. Durch diese Formulierung des systematischen Namens wird die katalysierte Reaktion spezifisch beschrieben, sodass der systematische Name in der Regel bereits eine Identifizierung und Klassifikation des entsprechenden Enzyms erlaubt. Somit ist der systematische Name vergleichsweise selbsterklärend und kann nach bestimmten Regeln durch die jeweiligen Entdecker eines Enzyms selbst geformt werden. Der systematische Name bildet häufig bei neu entdeckten Enzymen auch die Grundlage für den Trivialnamen. Dabei wird der systematische Name auf die wesentlichen Bestandteile verkürzt.

1.2.3 Klassifikationsschema nach IUPAC und IUBMB

Von der ersten Enzymkommission wurde ein numerisches Klassifikationssystem ausgearbeitet, das auf den sogenannten EC-(engl. Enzyme Commission) Nummern aufbaut. Formal setzt sich eine EC-Nummer aus dem Präfix EC-und vier Zahlen zusammen, die durch Punkte getrennt sind. Die erste Nummer zeigt, welcher Hauptklasse das entsprechende Enzym angehört. Die Hauptklassen sind gegliedert in Subklassen, die wiederum sind in Subsubklassen unterteilt sind. Dieser Gliederung folgend repräsentiert die zweite Nummer die Subklasse, während die dritte Nummer für die Subsubklasse steht. Die vierte und letzte Nummer ist eine Seriennummer des entsprechenden Enzyms innerhalb der Subsubklasse. Auf diese Weise wird die Klassifikation innerhalb des Nummerncodes immer spezifischer. Beispielsweise katalysiert die Glutamat-Dehydrogenase (systematischer Name: „L-glutamate:NAD+ oxidoreductase“) die Oxidation von Glutamat zu 2-Oxoglutarat unter der Beteiligung von H_2O und Abspaltung von Ammoniak.

$$\text{L-glutamate} + H_2O + \text{NAD+} \longrightarrow \text{2-oxoglutarate} + \text{NH3} + \text{NADH} + \text{H+}$$

Die EC-Nummer dieses Enzyms ist EC-1.4.1.2 und beginnt mit der Zahl Eins, da das Enzym der Enzymhauptklasse der Oxidoreduktasen angehört. Die nachfolgenden Zahlen des Nummerncodes werden in Abbildung 1 erläutert.

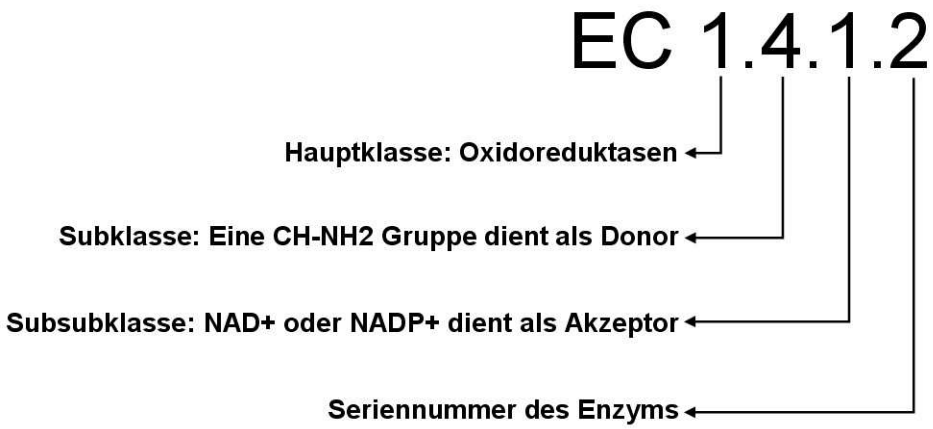


Abbildung 1 : Bedeutung des EC-Nummerncodes für das Enzym Glutamat-Dehydrogenase.

Neben der Hauptklasse der Oxidoreduktasen existieren 5 weitere Enzymhauptklassen. Einen Überblick über die Hauptklassen gibt Tabelle 1. Im Einzelnen unterscheiden sich die 6 Hauptklassen durch folgende Merkmale:

1. Oxidoreduktasen

Die Enzyme dieser Hauptklasse katalysieren Reduktions-Oxidations-Reaktionen (Redoxreaktionen). Bei diesem Reaktionstyp werden Elektronen von einem Reaktionspartner auf ein anderes Molekül oder Atom übertragen. Diese Elektronenübertragungs-Reaktionen lassen sich allgemein wie folgt beschreiben:

$$A^- + B \rightarrow A + B^-$$

Der Vorgang der Elektronenabgabe wird als Oxidation und das zugehörige Molekül als Elektronendonator bezeichnet. Die Reduktion beschreibt hingegen den Prozess der Elektronenaufnahme durch ein Akzeptormolekül. Der systematische Name enthält die Namen der Donator- und Akzeptormoleküle und basiert auf dem Schema Donor:Akzeptor Oxidoreduktase. Der Trivialname endet in der Regel auf Dehydrogenase, Reduktase oder Oxidase.

Von einigen Ausnahmen abgesehen hängt die zweite Stelle der EC-Nummer (Subklasse) davon ab, welche Molekülstrukturen Elektronen abgeben und somit als Elektronendonatoren fungieren.

Beispielsweise dient bei der ersten Subklasse eine Alkoholgruppe als Elektronendonor, während bei der zweiten Subklasse Aldehyd- oder Oxogruppen Elektronen abgeben.
Die dritte Stelle der EC-Nummer und somit die Subsubklasse hängt in der Regel von dem Elektronenakzeptormolekül ab. Die Subsubklasseneinteilung hängt entsprechend davon ab, ob NAD(P)+, Cytochrom, molekularer Sauerstoff oder andere Verbindungen Elektronen aufnehmen.

2. Transferasen

Transferasen katalysieren die Übertragung einer Molekülstruktur von einem sogenannten Donatormolekül zu einem anderen Akzeptormolekül. Die systematischen Namen werden nach dem Schema gebildet Donator:Akzeptor Gruppentransferase. Viele Transferase-Reaktionen lassen sich allgemein wie folgt formulieren:

$$A\text{-}C + B \rightarrow A + B\text{-}C$$

Hierbei wird die Molekülstruktur *C* von dem Donatormolekül auf das Akzeptormolekül *B* übertragen. Der Donator ist oft häufig ein Koenzym. Die Subklasseneinteilung orientiert sich an der Gruppe, die übertragen wird. So werden beispielsweise bei allen Reaktionen der Subklasse EC-2.1 einzelne Kohlenstoffeinheiten auf andere Moleküle transferiert, während bei Reaktionen der Subklasse EC-2.2 Aldehyd- oder Ketogruppen übertragen werden. Die übertragenden Gruppen werden für die Subsubklasseneinteilung feiner untergliedert und charakterisiert.

3. Hydrolasen

Das gemeinsame Merkmal der Hydrolasen ist die Spaltung einer Bindung unter Beteiligung eines Wassermoleküls. Allgemein lässt sich die Reaktionsgleichung wie folgt formulieren:

$$A\text{-}B + H_2O \rightarrow A\text{-}H + B\text{-}OH$$

Die Vertreter dieser Hauptklasse spalten Ether- und Esterbindungen, sowie glykosidische oder Peptidbindungen und eine Reihe weiterer Bindungen. Der systematische Name endet immer mit „Hydrolase“, während sich der Trivialname in vielen Fällen aus dem Substratnamen und dem Suffix –ase zusammensetzt. Die Subklassen werden anhand der gespaltenen Bindungen eingeteilt, während die dritte Stelle der EC-Nummer von den Eigenschaften des Substrats abhängt.

4. Lyasen

Lyasen ist ein Sammelbegriff für alle Enzyme, die Moleküle ohne die Beteiligung von Wassermolekülen spalten. Häufig entstehen bei Lyasereaktionen Doppelbindungen oder Ringstrukturen. Unter diese Hauptklasse fallen auch die Synthasen, die ohne Verbrauch von ATP

komplexe Molekülstrukturen aus einfachen Substrateinheiten herstellen. In diesem Fall bezieht sich die Bezeichnung „Lyase“ auf die Umkehrreaktion.
Die Subklasseneinteilung beruht auf dem Bindungstyp, der gespalten wird. Bei Subklasse EC-4.1 werden beispielsweise C-C-Bindungen gespalten. Die Reaktionen von Subklasse EC-4.2 basieren auf Spaltung einer C-O Bindung. Daneben gibt es 5 weitere Subklassen. Bei der Subsubklasseneinteilung wird der Molekülteil, der abgespalten wird, charakterisiert.

5. Isomerasen

Enzyme dieser Hauptklasse katalysieren geometrische oder strukturelle Änderungen von Molekülen, ohne dass sich dabei die Anzahl der Atome oder die Elementzusammensetzung des betreffenden Moleküls ändert. Unter Isomeren versteht man Moleküle mit gleicher Summenformel, aber unterschiedlicher Strukturformel.
Die verschiedenen Formen der Isomerie spiegeln sich in der Subklasseneinteilung wieder. Die Subklasse EC-5.1 enthält Racemasen, die eine Veränderung in der Stereo-Isomerie katalysieren. Dabei wird das Wasserstoffatom am chiralen Kohlenstoffatom eines Moleküls umgelagert, wodurch sich die optische Aktivität verändert (Optische Isomerie). Die Verknüpfungen innerhalb des Edukt- und Produktmoleküles ändern sich nicht. Beide Moleküle verhalten sich aber wie Bild und Spiegelbild (Spiegelbild-Isomerie) und sind durch Drehung nicht ineinander überführbar.
Subklasse EC-5.2 enthält cis-trans Isomerasen, welche die relative Stellung von zwei Substituenten innerhalb eines Moleküls verändern. In die Subklasse EC-5.3 fallen die intramolekularen Oxidoreduktasen, die Elektronentransferreaktionen innerhalb eines Moleküls katalysieren. Die Subklasse der intramolekularen Transferasen (EC-5.4) verschiebt funktionelle Gruppen innerhalb eines Moleküls an einen anderen Bindungsort (Mutasen). Schließlich gibt es noch die Subklassen der intramolekularen Lyasen und eine Reihe weiterer Isomerasen.
Die Subsubklasseneinteilung erfolgt auf Basis der Substrateigenschaften.

6. Ligasen

Ligasen verknüpfen Moleküle unter Verbrauch von Energieäquivalenten. In den meisten Fällen werden Nukleosidtriphosphate (NTP) verbraucht, wodurch sich folgende allgemeine Reaktionsgleichungen formulieren lassen:

$$Molekül\ 1 + Molekül\ 2 + NTP \rightarrow Molekül\ 1 - Molekül\ 2 + NDP + P$$

Manchmal wird Nukleosidtriphosphat auch zu Nukleosidmonophosphat und Pyrophosphat umgesetzt. In anderen Reaktionen dient NAD als Koenzym. Der systematische Name setzt sich aus den beiden Molekülen zusammen, die verbunden werden, in Kombination mit dem Suffix „Ligase". Früher wurde häufig für die Trivialnamen der Suffix „Synthetase" verwendet, bis diese Endung durch das Nomenklatur Komitee abgeschafft wurde. Die Trivialnamen der 6. Hauptklasse enden nun häufig auf „Ligase", „Synthase" oder „Carboxylase".

Hauptklasse	**Klassenname**	**Funktion**	**Vertreter**
EC-1.a.b.c	Oxidoreduktasen	Katalyse von Reduktions-Oxidations-Reaktionen	Monooxigenasen Dioxygenasen Oxidasen Dehydrogenasen Hydrogenasen Hydroxylasen
EC-2.a.b.c	Transferasen	Übertragung von Molekülstrukturen	Phosphotransferasen (enthalten Kinasen) Methyltransferasen Ethyltransferasen Aminotransferasen Polymerasen
EC-3.a.b.c	Hydrolasen	Hydrolytische Spaltung	Proteasen Peptidasen Nukleasen Phosphatasen Glykosidasen Esterasen
EC-4.a.b.c	Lyasen	Katalyse einer Molekülspaltung	Decarboxylasen Aldolasen Dehydratasen Synthasen
EC-5.a.b.c	Isomerasen	Katalyse geometrischer oder struktureller Veränderungen innerhalb eines Moleküls	Racemasen Epimerasen Tautomerasen Mutasen Cycloisomerasen Topoisomerasen
EC-6.a.b.c	Ligasen	Verknüpfen Moleküle unter Verbrauch von NTP	Synthasen Carboxylasen Cyclasen

Tabelle 1: Überblick über die Funktionen und Vertreter der 6 Enzymhauptklassen.

Die zweite Stelle der EC-Nummer und somit die Subklasseneinteilung hängt von den Bindungen ab, die in der Reaktion gebildet werden. Bei den Reaktionen der Subklasse EC-6.1 werden C-O-Bindungen geformt, während bei den nachfolgenden Subklassen C-S, C-N, C-C oder andere Bindungen gebildet werden. Fast alle Subklassen untergliedern sich in nur eine weitere Subsubklasse, abgesehen von Subklasse EC-6.3, in der C-N-Bindungen gebildet werden.

1.2.4 Nachteile des EC-Klassifikationssystems

Häufig genannte Kritikpunkte des EC-Klassifikationssystems sind, dass es nicht die evolutionären Beziehungen oder die dreidimensionale Struktur der Enzyme berücksichtigt. Eine Analyse der verwandtschaftlichen Beziehungen würde zu einer anderen Einteilung der Enzyme führen, ebenso wie eine Gliederung nach Proteindomänen.
Die internationale Enzymkommission stellte hingegen die katalysierte Reaktion der Enzyme in den Vordergrund. Die katalysierte chemische Reaktion erlaubt eine Darstellung der Wirkungs- und Substratspezifität, und damit eine Beschreibung der spezifischen Merkmale von Enzymen. Aus diesem Grund erwies sich das EC-Klassifikationssystem bisher als eine rationale Grundlage zur Enzymbenennung und Klassifizierung.
Dennoch ist die Zuordnung von neuen EC-Nummern hoch komplex und wird nicht von einzelnen Wissenschaftlern vorgenommen. Zuständig für die Bestimmung der EC-Nummer sind die Kommissionen der IUPAC und IUBMB. Zum einen gibt es die gemeinsame *Joint Commission on Biochemical Nomenclature* (JCBN), die beiden Organisationen zuzuordnen ist, zum anderen gibt es das *Nomenclature Committee of IUBMB* (NC-IUBMB). Die Vergabe einer neuen EC-Nummer erfordert die Absprache und das Einverständnis der beteiligten Kommissionen.
Ein anderes Problem neben der Komplexität sind Zuordnungen von Enzymen, die aus historischen Gründen oder physiologischen Aspekten gruppiert wurden. Daher unterliegt die Liste klassifizierter Enzyme einer ständigen Überarbeitung durch die Enzymkommissionen (WEBB EC, 1990). Neue Enzyme werden der Liste hinzugefügt und falsch klassifizierte Enzyme werden transferiert oder vollständig gelöscht.
Aus diesem Grund wäre ein System hilfreich, welches in der Lage ist, Reaktionen automatisch nach ähnlichen Kriterien zu bewerten, wie das EC-System. Hierdurch könnte die EC-Klassifikation unterstützt und vereinfacht werden.

1.3 Dugundji-Ugi-Modell

Das so genannte Dugundji-Ugi-Modell (Dugundji und Ugi, 1973) erlaubt eine Kodierung von chemischen oder enzymatischen Reaktionen. Ähnlich wie das EC-Klassifikationssystem betrachtet dieses Modell die Gesamtreaktion und dessen chemische Eigenschaften. Die Reaktionen werden mit Hilfe mathematischer Operatoren beschrieben, die als Reaktionsmatrizen bezeichnet werden und das Elektronentransfermuster von Reaktionen repräsentieren. Ursprünglich wurde das Dugundji-Ugi-Modell für die Beschreibung chemischer Reaktionen entwickelt.

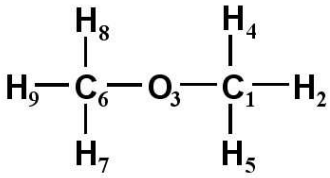
In diesem Modell werden Moleküle mit Hilfe so genannter Bindungselektronenmatrizen (BE-Matrizen) dargestellt. BE-Matrizen sind zweidimensionale Matrizen, die auch als Konnektivitätsmatrizen bezeichnet werden, da sie Aufschluss darüber geben, wie die Atome innerhalb eines Moleküls miteinander verbunden sind.

Abbildung 2 zeigt die BE-Matrix von Dimetylether. Die Atome des Moleküls sind entlang der Zeilen und Spalten aufgetragen, wobei jede Zeile und Spalte mit gleichem Index ein bestimmtes Atom des Moleküls repräsentiert. Die Werte innerhalb der Matrix geben das Bindungsverhältnis wieder. So bedeutet ein Matrixeintrag von Eins, dass die entsprechenden Atome über eine Einfachbindung miteinander verbunden sind. Ein Matrixeintrag von Zwei steht für eine Doppelbindung und ein Eintrag von Drei für eine Dreifachbindung. Die Werte entlang der Hauptdiagonalen korrespondieren mit den freien Valenzelektronen. Beispielsweise hat der Hauptdiagonaleneintrag des Sauerstoffatoms O3 einen Wert von 4, da Sauerstoffatome in der Regel 4 freie Valenzelektronen besitzen.

Der Index der Atome entlang der Zeilen und Spalten kann verändert werden, solange ein Atom die gleiche Zeilen- und Spaltenposition einnimmt. Von einer BE-Matrix existieren daher n! Permutationen, wobei n die Matrixdimension angibt. BE-Matrizen sind symmetrisch, können mehrere Moleküle kodieren und erlauben die Anwendung von Regeln der allgemeinen Algebra. Zur Darstellung von Reaktionen werden die Edukt- und Produktmoleküle mit Hilfe von BE-Matrizen kodiert. Die Reaktionen wurden durch die fundamentale Gleichung 1 ausgedrückt:

$$B + R = E \qquad (1)$$

BE-Matrix B repräsentiert die Eduktmoleküle und wird Eduktmatrix bezeichnet. Die Produktmoleküle werden durch die Produktmatrix E verkörpert. Matrix R ist die Reaktionsmatrix (R-Matrix), die als Transformationsoperator die Eduktmatrix in die Produktmatrix überführt. Reaktionsmatrizen geben Aufschluss darüber, welche Bindungen entstehen oder gespalten werden.



Dimethylether

BE-Matrix von Dimethylether

	C1	H2	O3	H4	H5	C6	H7	H8	H9
C1		1	1	1	1				
H2	1								
O3	1		4			1			
H4	1								
H5	1								
C6			1				1	1	1
H7						1			
H8						1			
H9						1			

Abbildung 2: Struktur und BE-Matrix von Dimethylether.

Positive Einträge stehen für Bindungen, die gebildet werden, wohingegen negative Werte Bindungen repräsentieren, die gespalten werden. Ein Reaktionsbeispiel mit den zughörigen BE-Matrizen und der R-Matrix zeigt Abbildung 3.

Es ist ebenso möglich, enzymatische Reaktionen mit Hilfe des Dugundji-Ugi-Modell zu beschreiben. Abbildung 4 zeigt als Beispiel die Reaktion der Aspartat Ammoniak-Lyase.
Die Summe aller Matrixeinträge in R ergibt in der Regel den Wert Null, denn die Anzahl der Elektronen im Reaktionssystem bleibt konstant. Eine Bindung, die auf der Eduktseite gespalten wird, führt zur Entstehung einer neuen Bindung auf der Produktseite oder äußert sich durch Zunahme der freien Valenzelektronen eines Atoms.
Das System setzt voraus, dass die Edukt- und Produktmatrix gleiche Dimension besitzen müssen. Diese Vorrausetzung lässt sich in der Regel leicht erfüllen, da die Anzahl der Atome im Verlauf der Reaktion konstant bleibt. Wesentlich schwieriger ist dagegen, die Vorraussetzung der Atomzuordnung zu erfüllen.

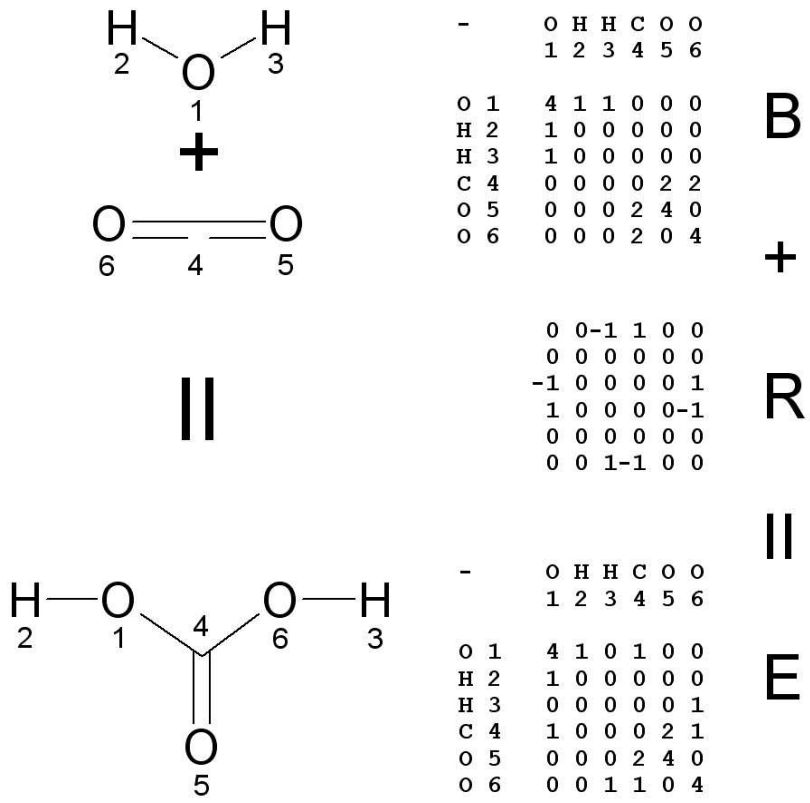


Abbildung 3: Chemische Reaktion von Kohlendioxid und Wasser zu Dihydrogencarbonat (Kohlensäure). Durch Addition der Reaktionsmatrix R zur Eduktmatrix B kann die Produktmatrix E errechnet werden.

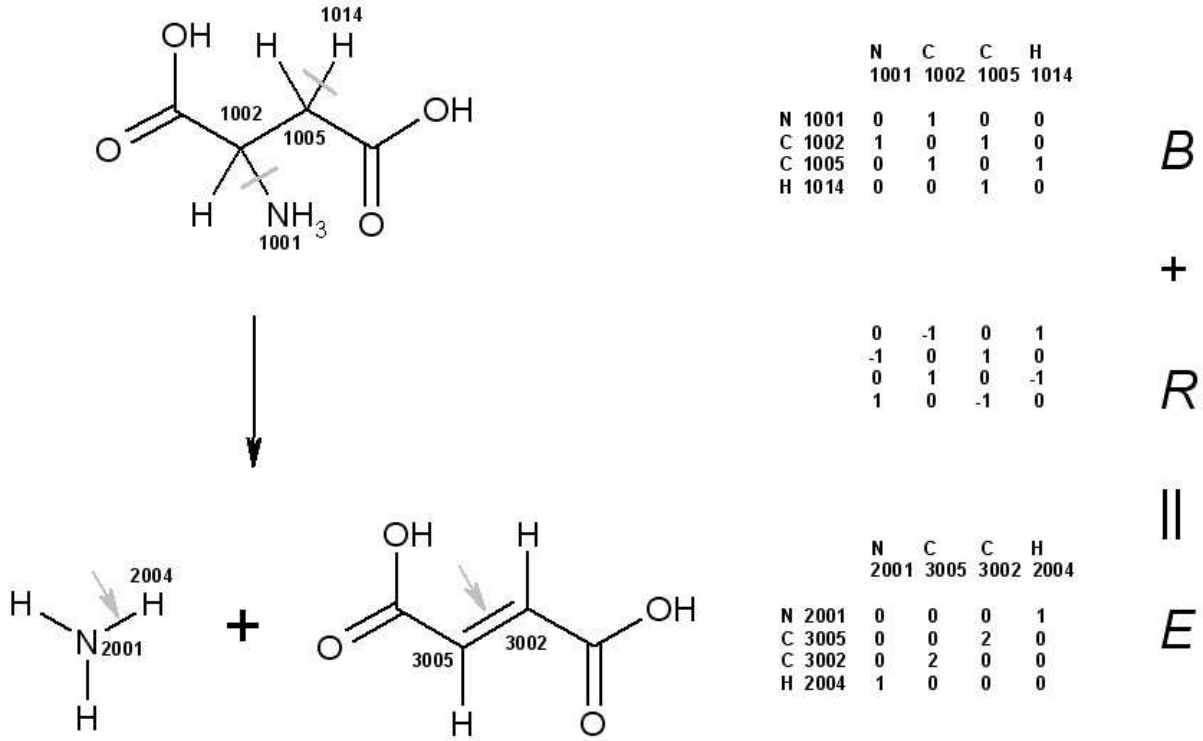


Abbildung 4: Reaktion der Aspartat Ammoniak-Lyase (EC-4.3.1.1), die L-Aspartat zu Ammoniak und Fumarsäure umsetzt. *B*, *R* und *E* beschränken sich auf den Reaktionskern.

1.3.1 Vorteile und Schwachpunkte des Dugundji-Ugi-Modells

Das Dugundji-Ugi-Modell ist ein objektives System, das sehr nüchtern das Muster des Elektronenaustausches einer Reaktion beschreibt. Über einfache arithmetische Rechenoperationen lässt sich der Reaktionskern mit den an der Reaktion beteiligten Atomen bestimmen. Als Grundlage für das System wurde ein mathematisches Konzept gewählt, das möglichst sachlich und unvoreingenommen die wichtigsten Eigenschaften einer Reaktion charakterisiert.

Aus diesem Grund wurde es als solide Basis für verschiedene anwendungsbezogene Programme verwendet. So wurden mit Hilfe des Dugundji-Ugi-Modells Methoden zur Reaktionsklassifizierung und Dokumentation, zur Generierung von Reaktionen und Aufklärung der beteiligten Mechanismen entwickelt (Ugi *et al.*, 1994). Ein Ansatz zur Klassifikation von chemischen Reaktionen auf der Grundlage von Elektronentransfermustern wurde von J. Brandt *et al.* (Brandt *et al.*, 1981, Brandt *et al.*, 1983) entwickelt. In diesen Arbeiten wurden Kanonisierungsregeln definiert zur Erzeugung einer eindeutigen R-Matrix aus der Menge von n! möglichen R-Matrizen, wobei n die Dimension der entsprechenden R-Matrix angibt. Die resultierende eindeutige R-Matrix wird als kanonische R-Matrix bezeichnet.

In anderen Anwendungen wurde das Dugundji-Ugi-Modell zur Generierung von Reaktionen benutzt. Die Programme IGOR (Bauer *et al.*, 1989) und RAIN (Fontain and Reitsam, 1991) sind hierfür zwei Beispiele. IGOR (interactive generation of organic reactions) benötigt eine gegebene R-Matrix und errechnet mögliche BE-Matrix Paare von Edukt- und Produktmatrizen. RAIN (reaction and intermediates networks) errechnet Produktmatrizen auf der Basis einer gegebenen B-Matrix. Beide Programme arbeiten aber nur mit Hilfe so genannter Übergangstabellen und einer Menge von formalen Beschränkungen, um eine kombinatorische Explosion von Möglichkeiten zu vermeiden (Ugi *et al.*, 1994).

Erstmals wurde das Dugundji-Ugi-Modell im Rahmen einer Diplomarbeit (Lüdge, 2002) auf biochemische Reaktionen angewandt. Für jede Subsubklasse wurde eine R-Matrix manuell erstellt und die R-Matrizen der Subsubklassen verglichen. In der neuesten Anwendung wurde das Dugundji-Ugi-Modell zur *de novo* Synthese von metabolischen Pfaden verwendet (Hatzimanikatis *et al.*, 2005). Diese Studie demonstriert, wie effektiv das Dugundji-Ugi-Modell zur Darstellung von biochemischen Reaktionen benutzt werden kann. Allerdings mussten die 250 Reaktionsmatrizen, die in dieser Studie verwendet wurden, mit hohem manuellen Aufwand erstellt werden.

Diese Beispiele machen deutlich, dass die Verwendung des Dugundji-Ugi-Modell fast immer nur mit einschränkenden Regeln oder mit hohem manuellen Aufwand möglich ist. Dieses erklärt, warum die Anzahl von Publikationen, die sich mit diesem System befassen, stark rückläufig ist.

1.3.2 Begrenzung des Dugundji-Ugi-Modells durch das Problem der Atomzuordnung

Eines der größten Hindernisse im Umgang mit dem Modell ist das Problem der Atomzuordnung. Beispielsweise kann die R-Matrix einer Reaktion leicht anhand von Gleichung 2 errechnet werden:

$$R = E - B \tag{2}$$

Die Matrixwerte ergeben sich aus $r_{ij} = e_{ij} - b_{ij}$, wobei i und j die Indices der Matrix repräsentieren. Die Reaktionsmatrix wird durch Substraktion der Eduktmatrix von der Produktmatrix errechnet. Wie in Abbildung 5 veranschaulicht, muss aber ein Atom innerhalb der Matrizen von B, E und R den gleichen Index einnehmen. Ist die Zuordnung der Eduktatome auf die Produktatome unbekannt, existiert kein nachvollziehbarer Reaktionsmechanismus und die Berechnung der R-Matrix ist unmöglich.

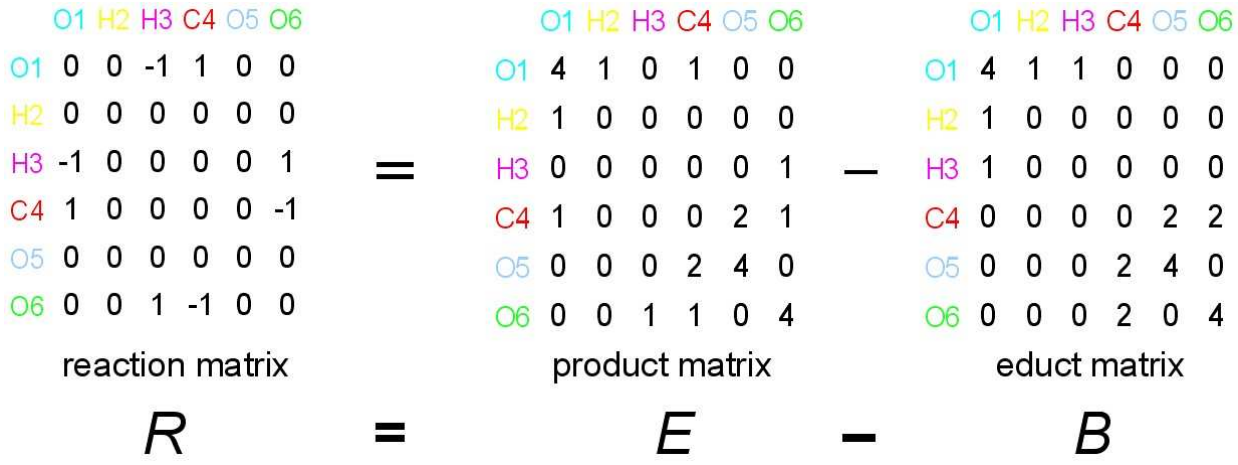


Abbildung 5: R-Matrix-Berechnung für die Reaktion von Kohlendioxid und Wasser zu Kohlensäure. Edukt- und Produktatome müssen innerhalb der Matrizen zugeordnet sein.

1.4 Zielsetzung

Die Einteilung der Enzyme beruht auf dem EC-Klassifikationssystem, das von Kommissionen der Organisationen IUPAC und IUBMB erarbeitet wurde. Es basiert im Wesentlichen auf den chemischen Eigenschaften der Reaktion, die von dem Enzym katalysiert wird, denn diese erlauben die Darstellung der Wirkungs- und Substratspezifität. Diese Eigenschaften sind die spezifischen

Merkmale von Enzymen und eignen sich daher in besonderer Weise zur Klassifikation und Enzymbenennung.
Dennoch ist die Zuordnung von neuen EC-Nummern hoch komplex. Zuständig für die Bestimmung der EC-Nummer sind Enzymkommissionen, wie die *Joint Commission on Biochemical Nomenclature* (JCBN) oder das *Nomenclature Committee of IUBMB* (NC-IUBMB), die eng zusammen arbeiten. Meinungsunterschiede, hinsichtlich der Eingliederung von Enzymen, haben vor allem in der Anfangszeit dazu geführt, dass Enzyme mehr nach historischen oder physiologischen Gesichtspunkten gruppiert wurden (WEBB EC, 1990). Die Enzymliste unterliegt daher einer ständigen Überarbeitung. So werden nicht nur neue Enzyme der Liste hinzugefügt, sondern auch Einträge verschoben oder vollständig entfernt.

Die Komplexität des EC-Klassifikationssystems und die fragwürdige Einordnung einiger Enzymgruppen war eine Motivation für die vorliegende Arbeit, ein neues automatisiertes Verfahren für die Klassifikation von enzymatischen Reaktionen einzusetzen. Dem Verfahren liegt das so genannte Dugundji-Ugi-Modell (Dugundji und Ugi, 1973) zugrunde. Ähnlich dem EC-Klassifikationssystem betrachtet dieses System die chemischen Eigenschaften der Gesamtreaktion. Reaktionen werden durch mathematische Operatoren beschrieben, die als Reaktionsmatrizen (R-Matrizen) bezeichnet werden und die Elektronentransfermuster einer Reaktion kodieren. R-Matrizen enthalten die Information, welche Bindungen gespalten werden oder entstehen und welche Atome an der Reaktion beteiligt sind. Das Errechnen der R-Matrizen erfordert allerdings eine Atomzuordnung der Eduktatome auf die Produktatome. Diese Zuordnung wurde in der Vergangenheit immer mit hohem manuellem Aufwand erstellt.

Somit bestand das primäre Ziel dieser Arbeit in der Entwicklung eines automatisierten Verfahrens, das die R-Matrizen nur anhand der Edukt- und Produktmoleküle errechnet und ohne manuelle Unterstützung auskommt. Außerdem sollte ein geeignetes Vergleichssystem für die R-Matrizen entwickelt werden, dass einen effizienten Vergleich und die Gruppierung der R-Matrizen ermöglicht. Schließlich ging es darum, abzuschätzen, inwiefern das System eine sinnvolle Gruppierung von enzymatischen Reaktionen ermöglicht oder ob es zur Unterstützung des EC-Klassifikationssystems hilfreich sein kann.

1.5 Vergleichende Arbeiten der Computergestützten Analyse von Reaktionen

Die KEGG (Kyoto Encyclopedia of Genes and Genomes)-Datenbank (Kanehisa *et al.*, 2008) entwickelte ein Programm zur automatischen Vorhersage von EC-Nummern, das 2004 publiziert wurde (Kotera *et al.*, 2004). In diesem Verfahren wird jede Reaktionsformel in Sets zerlegt, die jeweils aus den korrespondierenden Edukt- und Produktmolekülen bestehen. Die Zuordnung dieser Reaktionspaare erfolgt manuell. In einem zweiten Schritt wird das Strukturvergleichsprogramm SIMCOMP (Hattori *et al.*, 2003). dazu verwendet, die Moleküle der Reaktionspaare miteinander zu vergleichen. Innerhalb dieser Methode erhält jedes Atom ein spezifisches Label je nachdem, welcher funktionellen Gruppe es angehört. Der Algorithmus versucht die Reaktionsstellen zu identifizieren und gliedert die Atome in drei Klassen, je nachdem, ob sie unmittelbar an das Reaktionszentrum angrenzen, sich in der näheren Umgebung zu diesem befinden oder weiter entfernt sind. Die Veränderungen werden jeweils in Form von numerischen Codes angegeben, die gesammelt und zu sogenannten „Reaction Classification“ (RC) Nummern kombiniert werden. Wie KEGG berichtet, führt diese Methode in vielen Fällen zu suboptimalen Ergebnissen (Oh *et al*., 2007). Aus diesem Grund wurden die Molekülüberlagerungen, die durch das Programm SIMCOMP errechnet wurden, in den letzten Jahren manuell, Reaktion für Reaktion, überarbeitet.

Die Fehler liegen zum einen in dem Programm SIMCOMP begründet, das für den Molekülvergleich einen Cliquenalgorithmus verwendet. Dieser benutzt Heuristiken, die mit zunehmender Molekülgröße zu falschen Ausgaben führen können. Zum anderen genügt es nicht die korrespondierenden Edukt- und Produktmoleküle paarweise zu vergleichen, da es häufig zu Überlagerungen von Strukturen kommt.

Um solche Effekte zu umgehen, wird in der vorliegenden Arbeit ein Rankingsystem benutzt, das überlagernde Strukturen erkennt und die entsprechenden Atome nur einmal eindeutig zuordnet (vgl. 2.2.2). Anstelle des SIMCOMP Programms wurde ein neuer *c*-MCS-Algorithmus entwickelt (vgl. 2.1.6), der auch große Moleküle vergleichen kann und zudem flexibel genug ist, um auch komplexe Veränderungen in der Struktur nachzuvollziehen. Das von KEGG entwickelte Verfahren beruht nicht auf dem Dugundji-Ugi-Modell und verwendet keine Elektronentransfermuster.

2. Methoden

Die Entwicklung eines automatisierten Verfahrens zur Berechnung von R-Matrizen erfordert zuerst eine präzise Atomzuordnung der Eduktatome auf die Produktatome. Kern dieses Verfahrens ist ein Algorithmus, der die größte gemeinsame Substruktur von Molekülen bestimmt. Darüber hinaus war die Entwicklung zahlreicher weiterer Module erforderlich, die in das Programm integriert wurden. Einen Überblick über die Gesamtstruktur des Programms gibt Abbildung 6.

Zusammengefasst folgt das Verfahren folgenden Schritten. Zunächst wird der Algorithmus zur Bestimmung der größten gemeinsamen Substruktur zum Vergleich der Eduktmoleküle mit den Produktmolekülen verwendet. Zwei Moleküle besitzen häufig mehr als eine größte gemeinsame Substruktur, sodass der Vergleich zu einem Set von größten gemeinsamen Substrukturen führt.

Im nächsten Programmabschnitt wird eine Atomzuordnung für die Reaktion generiert, indem die größten gemeinsamen Substrukturen der verschiedenen Sets in allen möglichen Zusammenstellungen kombiniert werden. Dieser komplexe Schritt erfordert den Einsatz eines Rankingsystems und wird in Abschnitt 2.2.2 genauer beschrieben.

Als Eingabe werden dem Programm sogenannte Molfiles übergeben, die jeweils für ein Molekül kodieren. Molfiles enthalten nur in Ausnahmefällen Informationen über die Wasserstoffatome des Moleküls. Daher werden die Wasserstoffatome mit Hilfe der anderen Atome des Molekülrückrats ergänzt und ebenfalls zugeordnet.

In einigen Fällen wurden spezielle Strategien in das Programm integriert. Zum Beispiel können aromatische Ringe zu Pseudomustern in den R-Matrizen führen. Dieser Effekt hängt mit der Molfilekodierung der Moleküle zusammen und wird in Abschnitt 2.2.7 genauer erläutert. Es wurde ein Modul programmiert und integriert, das aromatische Ringe auffindet und diese Effekte ausschließt.

Ein anderes Beispiel ist der Vergleich von kleinen Molekülen mit großen Molekülen. Der Vergleich führt häufig zu einer hohen Anzahl von größten gemeinsamen Substrukturen. Diese Zuordnungen sind häufig nicht signifikant und besitzen aufgrund ihrer geringen Größe innerhalb des Rankingsystems eine niedrige Priorität. Es werden daher kleine Moleküle zunächst nicht berücksichtigt und später mit Hilfe einer anderen Variante des Bron-Kerbosch-Algorithmus zugeordnet.

Wenn die Atomzuordnung abgeschlossen ist, können schließlich die Edukt- und Produktmatrizen generiert werden und die Berechnung der R-Matrix kann anhand von Gleichung 2 (vgl. 1.3.2) erfolgen. In einigen Fällen existieren mehrere gleichwertige atomare Zuordnungen und somit R-Matrizen für eine Reaktion, wie in Abschnitt 2.2 beschrieben wird.

Zuletzt erfolgt der Schritt der so genannten Kanonisierung. Dieses Verfahren ermöglicht einen besseren Vergleich und die Erzeugung einer eindeutigen R-Matrix. Problematisch für den Vergleich von R-Matrizen ist die hohe Anzahl von n! möglichen Permutationen, wobei n die Matrixdimension repräsentiert. Da es je nach Matrixdimension zu zeitaufwändig ist, alle n! möglichen Permutionen einer R-Matrix zu generieren, wird nach bestimmten Regeln die Ausgangsmatrix umgeformt, sodass eine eindeutige, kanonische Matrix entsteht.

Alle Module wurden in C++ implementiert. Da rechenzeitintensive Operationen innerhalb der R-Matrix-Berechnung vorhersehbar waren, schieden interpretierende Sprachen, wie Python, Perl oder Java aus. Daneben wurden kleine Programme in Python ausgeführt, die Datenbankabfragen oder Progammaufrufe vornehmen. Die Verfahren und Algorithmen, die den verschiedenen Modulen zugrundeliegen, werden in den folgenden Abschnitten genau erläutert.

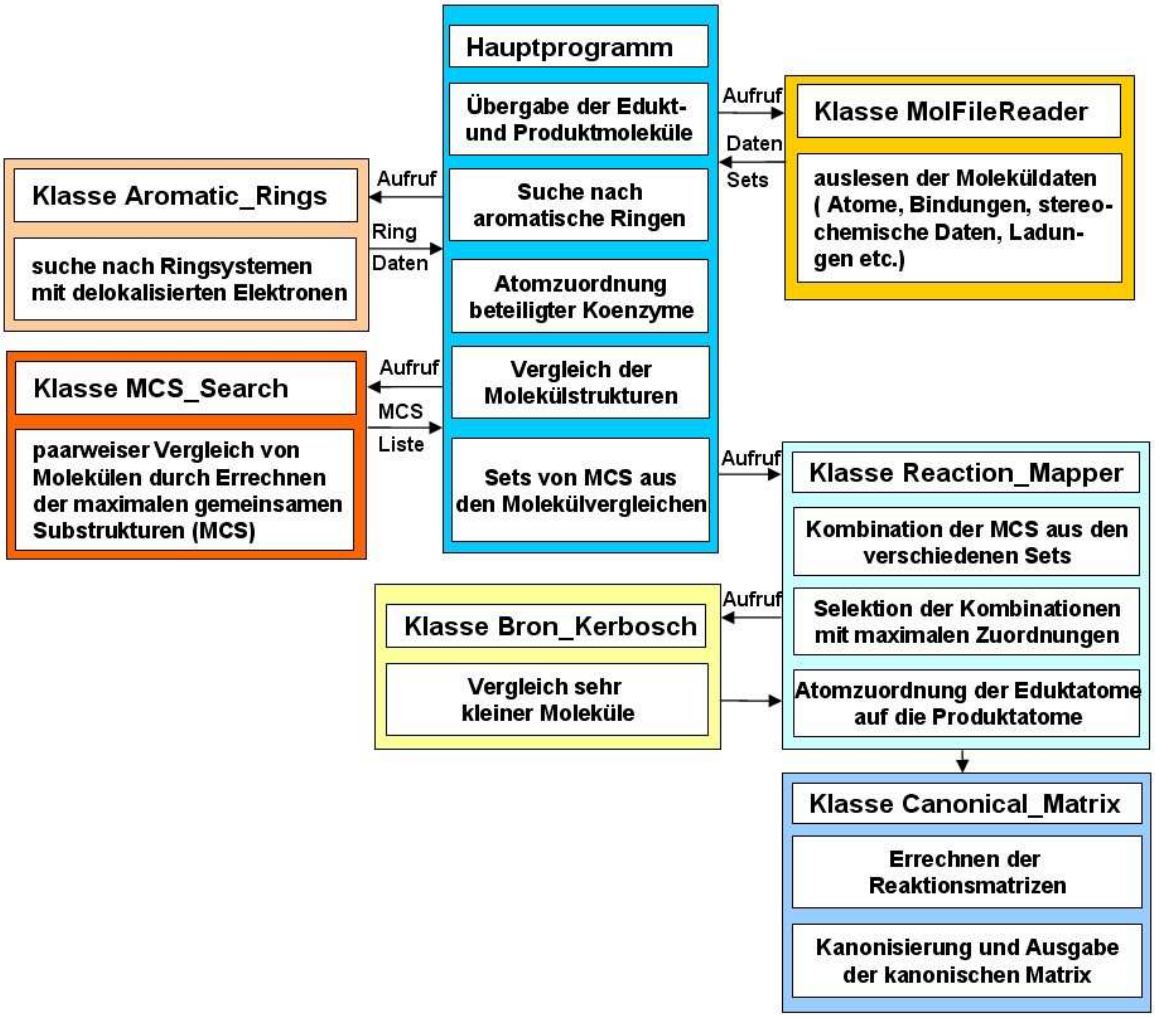


Abbildung 6: Gesamtstruktur des Programms

2.1 MCS Algorithmen

Eine wichtige Vorraussetzung für die R-Matrix-Berechnung ist eine genaue Atomzuordnung der Edukt- und Produktatome. Eine solide Grundlage für eine korrekte Atomzuordnung sind Methoden, die Molekülstrukturen vergleichen können. Der grundlegende Gedanke ist hierbei, dass ein Eduktatom einem Produktatom zugeordnet werden kann, wenn das Atom sowohl vor als auch nach der Reaktion der gleichen Molekülstruktur angehört.
Die effezientesten Algorithmen zum Vergleich von Molekülen basieren auf mathematischen Verfahren. In der Graphentheorie wurden Algorithmen zur Bestimmung der maximalen gemeinsamen Subgraphen von zwei Graphen entwickelt. Diese Algorithmen werden in der Literatur häufig als MCS Algorithmen bezeichnet, wobei sich die Abkürzung „MCS“ aus den Anfangsbuchstaben der englischen Bezeichnung „Maximal Common Subgraph“ ableitet.

2.1.1 Graphenbasierte Darstellung von Molekülen

Moleküle lassen sich als Graphen repräsentieren. Ein Graph *G* ist ein Gebilde von Knoten *V* (*V* bedeutet „*Vertex*“, deutsch Knoten), die über Kanten *E* (*E* bedeutet „*Edge*“, deutsch Kante) miteinander verdunden sind. Definiert ist ein Graph als ein Paar zweier endlicher Mengen:

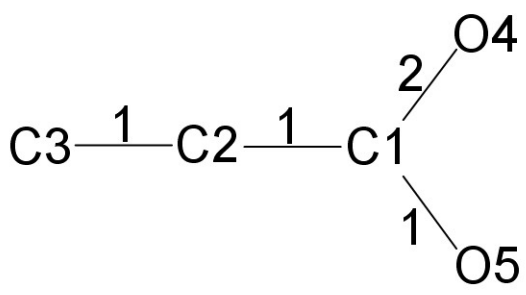
$$\text{Graph } G = (V,E)$$

Zudem besitzt jeder Graph eine auf der Kantenmenge *E* definierte Abbildung Ψ:

$$\Psi : E \rightarrow V^2$$

Ein Beispiel für die graphenbasierte Darstellung eines Moleküls zeigt Abbildung 7. Die Kohlenstoff- und Sauerstoffatome von Propansäure repräsentieren die Knoten des Graphen. Auf die Darstellung der Wasserstoffatome wurde verzichtet. Die Kanten des Graphen werden hingegen durch die Bindungen beschrieben. Die Kanten sind ungerichtet, aber gewichtet. Die Gewichtung der Kanten gibt Aufschluss über die Anzahl der Bindungen zwischen 2 Atomen. Die Atome sind weitgehend über eine Einfachbindung verbunden, abgesehen von der Bindung zwischen dem Kohlenstoffatom C1 und dem Sauerstoffatom O4, die über eine Doppelbindung verbunden sind.
Für die vorliegende Arbeit wurden sogenannte Molfiles verwendet, in denen Moleküle in Form des MDL-Formats kodiert sind. Dieses Speicherformat bedient sich ebenfalls einer graphenbasierten Darstellung von Molekülen. Aus diesem Grund lassen sich die Sets von *V* und *E* eines Moleküls

unmittelbar aus den Dateien auslesen, sodass dieses Molekül als Graph weiterverarbeitet werden kann.



Knotenset *V*: C1, C2, C3, O4, O5

Kantenset *E*:

Kanten	Label
C1, C2	1
C2, C3	1
C1, O4	2
C1, O5	1

Abbildung 7: Darstellung von Propansäure als Graph. Die Atome repräsentieren die Knoten und die Bindungen die Kanten des Graphen.

2.1.2 Bestimmung der maximalen gemeinsamen Substruktur

MCS Algorithmen erlauben den paarweisen Vergleich von Graphen und die Bestimmung des maximalen gemeinsamen Subgraphen. Das MCS-Problem kann wie folgt definiert werden: Gegeben seien 2 Graphen $G_1 = (V_1,E_1)$ und $G_2 = (V_2,E_2)$. Dann werden maximale gemeinsame Subgraphen definiert als $H_i = (V,E)$, wobei H isomorph zu Subgraph $G'_1 = (V'_1,E'_1)$ von Graph G_1 und $G'_2 = (V'_2,E'_2)$ von Graph G_2 ist.

Da Moleküle ebenfalls Graphen repräsentieren, ist die Suche nach dem maximalen gemeinsamen Subgraphen von zwei Graphen äquivalent zu der Suche nach der maximalen gemeinsamen Substruktur von zwei Molekülen. Aus diesem Grund werden MCS Algorithmen seit vielen Jahren zum Vergleich von Molekülen eingesetzt (Hattori *et al.*, 2003, Marialke *et al*., 2007, Raymond *et al*., 2002, García *et al*., 2004, Durand *et al*., 1999, McGregor *et al*., 1982, Bayada *et al*., 1992).
Der Vorteil bei der Verwendung von MCS Algorithmen ist eine sehr genaue Überlagerung der Molekülstrukturen. Andererseits ist die MCS-Suche ein NP-vollständiges Problem („NP“ bedeutet „nichtdeterministisch polynomielle Zeit“). Für die Probleme dieser Komplexitätsklasse konnte bisher kein Algorithmus gefunden werden, der ihr Wortproblem in polynomieller Zeit lösen kann. Die Laufzeit ist exponentiell. Während die Suche für kleine Eingabegraphen in akzeptabler Zeit gelöst werden kann, nimmt sie mit der Größe der Eingabegraphen exponentiell zu. Daher wird in Bezug auf die Größe der Eingabegraphen sehr schnell eine Grenze erreicht, ab der eine MCS-Berechnung nicht mehr effizient ist.

Aus diesem Grund bedienen sich MCS Algorithmen heuristischer Methoden. Bei den Molekülvergleichalgorithmen kann die Laufzeit verbessert werden, indem zum Beispiel Doppelbindungen, die atomare Umgebung oder die dreidimensionale Struktur berücksichtigt werden. Diese zusätzlichen Daten ermöglichen eine bessere Unterscheidung der Atome als Eingabeknoten. Durch diese Einschränkungen wird einerseits die Laufzeit verbessert, doch andererseits führt dies zu einer stärkeren Begrenzung der MCS Größe, je unähnlicher die Eingabemoleküle werden. Letzteres betrifft besonders die Edukt- und Produktmoleküle von Reaktionen. Im Verlauf einer Reaktion können Einfachbindungen zu Doppelbindungen werden, die atomare Umgebung der Atome kann sich ändern und es kann zu Veränderungen der dreidimensionalen Struktur kommen. Somit gibt es einen Konflikt zwischen Laufzeitverbesserung und der Flexibilität des Algorithmus.
Für die Betrachtung enzymatischer Reaktionen wurde ein flexibler Algorithmus benötigt, der auch in der Lage ist, komplexe Veränderungen der Molekülstruktur nachzuvollziehen. Andererseits muss auch ein Vergleich größerer Metabolite möglich sein. Da ein entsprechender Algorithmus nicht verfügbar war, wurde innerhalb dieses Projektes ein neuer Algorithmus entwickelt. Dieser neue MCS-Algorithmus basiert auf zwei anderen gängigen MCS Algorithmen und kombiniert deren verschiedene positiven Eigenschaften. Dabei handelt es sich um eine schnelle Variante des Bron-Kerbosch-Algorithmus und den McGregor-Algorithmus, der für die Betrachtung chemischer Reaktionen entwickelt wurde.

2.1.3 Bron-Kerbosch-Algorithmus

Der bekannteste und am weitesten verbreitete MCS-Algorithmus ist der Bron-Kerbosch-Algorithmus (Bron und Kerbosch, 1973). Dieser Algorithmus gehört zu den Verfahren, die das MCS-Problem lösen, indem sie es in das Cliquen-Problem überführen. Dabei lässt sich das Problem in zwei Schritten lösen. Zunächst wird aus den beiden Eingabegraphen ein übergeordneter Produktgraph (Hartnell *et al.*, 1998) generiert. In einem zweiten Schritt erfolgt eine Cliquensuche in dem Produktgraphen. Die Generierung des Produktgraphen, die Definition einer Clique und die Suche nach den maximalen Cliquen werden in den nächsten Abschnitten näher erläutert.

Generierung des Produktgraphen:

Der Produktgraph gibt Aufschluss über die Ähnlichkeiten zwischen einem Graphen G_1 und einem Graphen G_2. Zugrunde liegt die Annahme, dass einige Knoten und Kanten eines Graphen G_1 kompatibel sind zu einigen Knoten und Kanten eines zweiten Graphen G_2. Alle diese Kompatibilitäten werden in einem neuen Graphen abgespeichert, indem das Produkt G_1 x G_2 aus den beiden Eingangsgraphen gebildet wird. Der neue Graph wird als Kompatibilitätsgraph oder Produktgraph bezeichnet. Dieser Graph wird anhand bestimmter Regeln aufgebaut.

Je nachdem, ob nach induzierten gemeinsamen Subgraphen oder nach gemeinsamen Subgraphen gesucht werden soll, wird entweder ein „Knoten-Produkt-Graph“ H_v oder ein „Kanten Produkt Graph“ H_e generiert. Im der vorliegenden Arbeit wurde ein "Knoten-Produkt-Graph" verwendet, da dieser direkt Rückschlüsse auf die atomaren Zuordnungen erlaubt.

Zunächst wird das Knotenset V_H des Produktgraphen H_v gebildet auf Grundlage des kartesischen Produktes $V_H = V(G_1)$ x $V(G_2)$. Hierbei werden aus den Knotenmengen der beiden Eingangsgaphen geordnete Paare (u,u') gebildet, wobei $u \in V(G_1)$ und $u' \in V(G_2)$ ist, sowie u und u' gleiche Label besitzen. Abbildung 8 zeigt ein Beispiel, wie die Knoten des Produktgraphen gebildet werden für den Vergleich von Propanol und Isopropanol. Alle Atome eines Elementes von Propanol werden mit allen Atomen des gleichen Elementes von Isopropanol gepaart. Jedes Kohlenstoffatom von Propanol wird mit jedem Kohlenstoffatom von Isopropanol zu einem Knoten gepaart. Da jedes Molekül drei Kohlenstoffatome besitzt, entstehen auf diese Weise 9 Knoten. Die Moleküle besitzen weiterhin je ein Sauerstoffatom, wodurch ein weiterer Knoten entsteht.

Nach Bildung der Knoten folgt die Definition der Kanten. Bei Produktgraphen können zahlreiche Definitionen zur Festlegung der Kanten verwendet werden, die zu unterschiedlichen Graphen führen (Levi, 1972). Zwei Knoten (u,u') und (v,v') des Produktgraphen werden durch eine Kante verbunden, wenn $u \neq v$ und $u' \neq v'$ ist, und wenn gilt:

A) u ist mit v in G_1 und u' ist mit v' in G_2 über eine Kante verbunden, oder
B) u ist mit v in G_1 und u' ist mit v' in G_2 nicht verbunden

Die Kanten in dem Produktgraph werden als *c*-Kanten („*c*“ bedeutet „connected“, deutsch „zusammenhängend“) bezeichnet, wenn Bedingung A erfüllt wird. Bei Erfüllung von Bedingung B, wird die Kante als *d*-Kante bezeichnet („*d*“ bedeutet „disconnected“, deutsch „getrennt“). In Abbildung 9 sind die *c*-Kanten innerhalb des Produktgraphen des Propanol - Isopropanol - Vergleiches grau gekennzeichnet. Die *d*-Kanten sind hingegen schwarz.

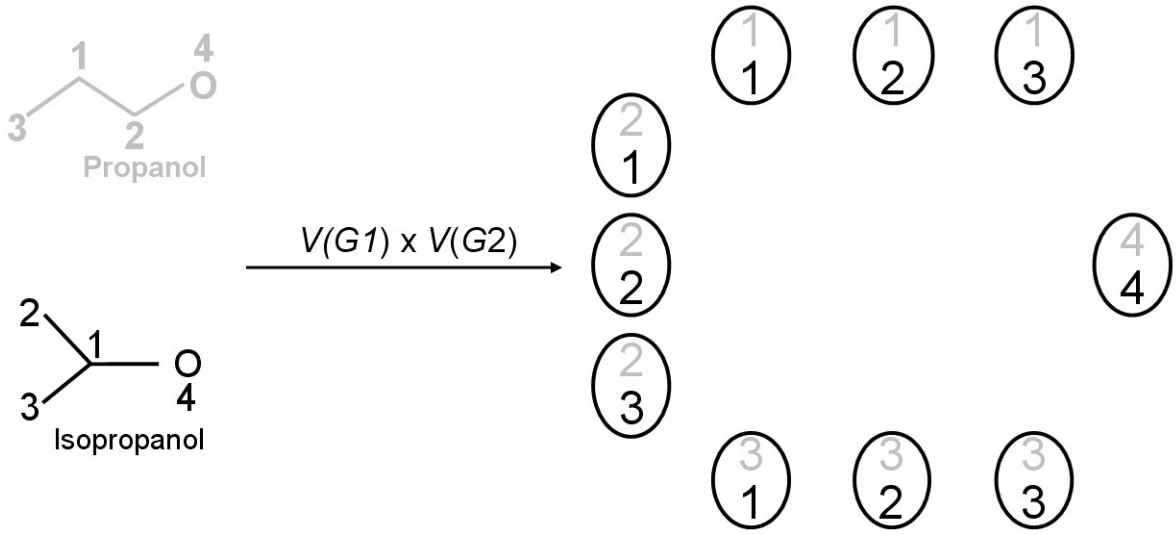


Abbildung 8: Bildung des Knotensets V_H des Produktgraphen H_v für den Vergleich von Propanol und Isopropanol. Die ellipsenförmigen Kreise beinhalten je ein Atompaar der beiden Moleküle. Die grauen Nummern stehen für die Atome des Propanol, während die schwarzen Nummern die Atome des Isopropanol repräsentieren.

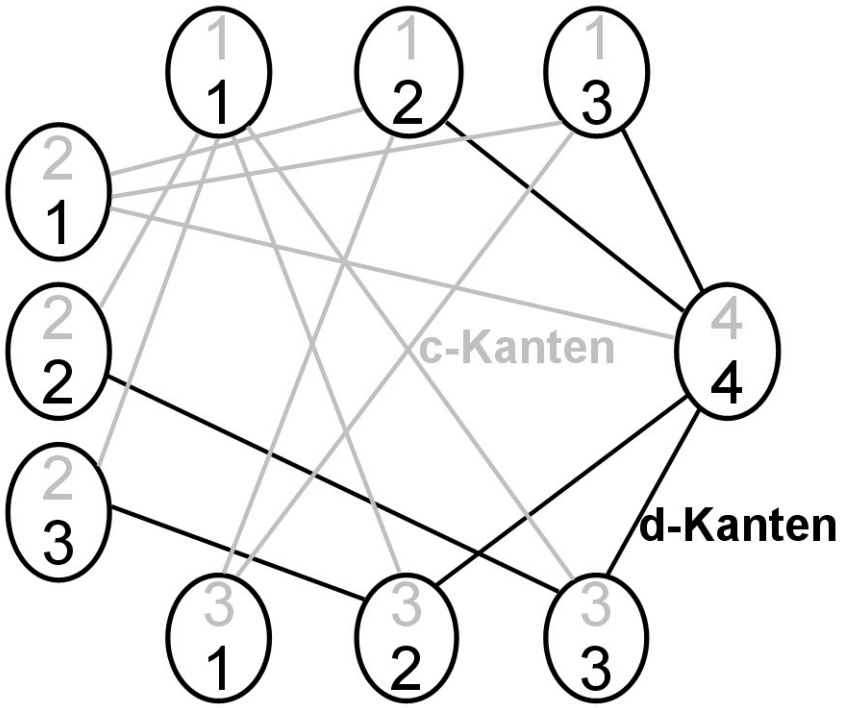


Abbildung 9: Bildung der Kanten für den Produktgraphen H_v für den Vergleich der Moleküle Propanol und Isopropanol. Die *c*-Kanten sind grau und die *d*-Kanten schwarz eingezeichnet.

Cliquensuche innerhalb des Produktgraphen:

Eine Clique ist eine spezielle Teilmenge von Knoten eines ungerichteten Graphen. Diese Knotenteilmenge repräsentiert einen vollständigen Teilgraphen. Dies bedeutet, dass jeder Knoten mit jedem anderen Knoten innerhalb der Clique über eine Kante verbunden ist. Eine Clique ist dann maximal, wenn kein weiterer Knoten des Graphen zur Clique hinzugefügt werden kann.

Das so genannte Cliquenproblem beschäftigt sich mit der Frage, ob ein Graph Cliquen der Größe k enthält, wobei k eine natürliche Zahl repräsentiert. Häufig geht es darum, die größten Cliquen eines Graphen zu ermitteln. Das Cliquenproblem gehört ebenso zu den NP-vollständigen Problemen und läßt sich nicht für beliebig große Graphen lösen. Dennoch bietet die Suche nach den maximalen Cliquen eine elegante Lösung für das MCS-Problem. Wie bei Levi (1972) beschrieben, korrespondieren Cliquen in dem Produktgraphen *H* mit gemeinsamen Subgraphen der beiden Eingangsgraphen G_1 und G_2. Dementsprechend korrespondieren maximale Cliquen in *H* mit maximalen gemeinsamen Subgraphen von G1 und G2.

Abbildung 10 zeigt die maximalen Cliquen in dem Produktgraphen H_v aus dem Vergeich von Propanol mit Isopropanol. In dem Produktgraphen gibt es 4 maximale Cliquen, die in der Abbildung jeweils durch graue Kanten gekennzeichnet sind. Jede maximale Clique umfasst 3 Knoten und korrespondiert mit einer MCS von Propanol und Isopropanol.

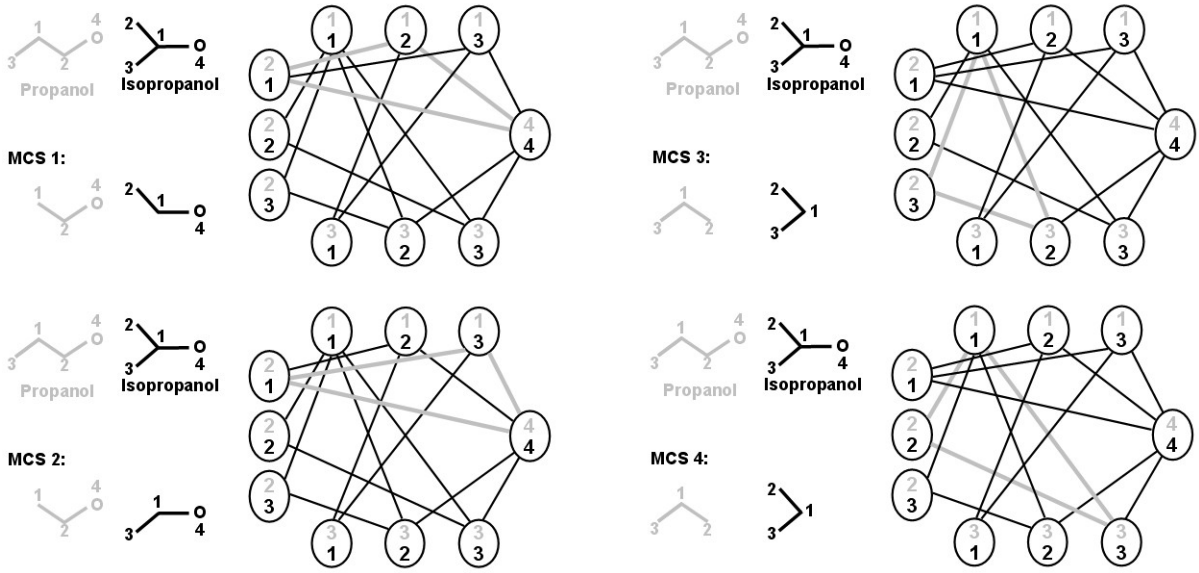


Abbildung 10: Maximale Cliquen in dem Produktgraphen von Propanol und Isopropanol korrespondieren mit maximalen gemeinsamen Substrukturen der beiden Moleküle.

Für die Suche nach maximalen Cliquen in Graphen wurden verschiedene Algorithmen mit exponentieller Laufzeit entwickelt. Der Bron-Kerbosch-Algorithmus löst dieses Problem über ein Rücksetzverfahren oder englisch Backtracking (engl. Rückverfolgung).
Dieses Verfahren beruht auf dem Prinzip Versuch und Irrtum und läßt sich häufig in Form eines Suchbaumes darstellen. Ausgehend von einem Startpunkt werden Teillösungen ausgetestet, die schrittweise weiter ausgebaut werden. Abbruchbedingungen für einen Suchweg innerhalb des Suchbaumes werden definiert. Diese Bedingungen machen kenntlich, dass entweder eine neue Gesamtlösung gefunden wurde oder ein Weg nicht mehr zu einer besseren Lösung führen kann. Wird eine neue Gesamtlösung gefunden oder wird ersichtlich, dass der Weg zu keiner besseren Lösung führt, erfolgt das Rücksetzen. Dabei wird zur letzten Verzweigung zurückgegangen und der nächste Weg des Suchbaumes getestet. Auf diese Weise werden systematisch alle Lösungsmöglichkeiten geprüft. Backtracking-Algorithmen erlauben somit das Auffinden einer Lösung, falls es eine Lösung gibt, oder eine Lösung kann definitiv ausgeschlossen werden.
Der Bron-Kerbosch-Algorithmus läßt sich mittels Rekursion implementieren. Rekursion ist eine elegante Strategie zur Lösung von komplexen Problemen. Der Pseudocode des Algorithmus wird auf der nächsten Seite wiedergegeben (Algorithmus 1). Einmal aufgerufen, ruft sich die Funktion "Errechne_Cliquen" selbst auf. Dabei entspricht jeder Aufruf einer Verzweigung innerhalb eines Suchbaumes, der durch den rekursiven Funktionsaufruf aufgebaut und abgesucht wird, bis alle maximalen Cliquen gefunden wurden. Die Knoten des Graphen werden durch die Funktion in die Gruppen *C*, *P* ist *S* eingeteilt.

Set *C*: beinhaltet Knoten, die bereits betrachtet und zum gegenwärtigen Cliquenset hinzugefügt wurden

Set *P*: Knoten, die noch nicht betrachtet wurden, aber zum gegenwärtigen Cliquenset (Set *C*) hinzugefügt werden könnten, da sie mit allen bisherigen Knoten aus Set *C* über eine Kante verbunden sind

Set *S*: Knoten, deren Cliquen bereits betrachtet wurden und von denen bekannt ist, dass sie nicht zu Set *C* hinzugefügt werden können

Diese Sets verändern sich bei jedem Schritt innerhalb des Backtracking-Baumes in Abhängigkeit von dem betrachteten Suchknoten u_i. Zunächst befinden sich alle Knoten in Set *P*, während die Sets *C* und *S* leer sind. Jeder Knoten aus *P* dient als Startpunkt für die Backtracking-Suche. Wird ein Knoten u_i aus *P* als Startpunkt gewählt, ist er der erste Knoten des Lösungssets *C*. Das Set *P* reduziert sich bei jedem Schritt auf die Knoten, die mit dem Suchknoten u_i benachbart sind. Die

verbleibenden Knoten in Set *P* dienen als Verzweigungen in dem folgenden Suchschritt. Bei den Knoten von Set *S* handelt es sich um Nachbarknoten des jeweiligen Suchknotens, die auf der jeweiligen Suchebene bereits betrachtet wurden. Somit wurden auch Lösungen, die mit diesen Knoten in Verbindung stehen, bereits betrachtet. Die Knoten aus *S* können demnach in Verbindung mit dem jeweiligen Suchknoten u_i zu keiner neuen Lösung führen. Voraussetzung für das Auffinden einer neuen Lösung ist somit, dass die Sets *S* und *P* leer sind.

Errechne_Cliquen(*C*,*P*,*S*)

> errechne alle Cliquen eines beliebigen Graphen *G*
C: Knoten der gegenwärtigen Clique
P: Knoten, die noch zu Set *C* hinzugefügt werden können
S: Knoten, die nicht zu Set *C* hinzugefügt werden können
N[*u*]: Set von Nachbarknoten von Knoten *u* in *G*

```
01    P is the Set {u1,...,uk}
02    if  P=∅ and S=∅
03          then report Clique;
04    else  for i ← 1 to k
05          do
06                P ← P \ {ui,};
07                P' ← P;
08                S' ← S;
09                N ← {v ∈ V | {ui,v} ∈ E}
10                Errechne_Cliquen(C ∪ {ui}, P'∩N, S'∩N);
11                S ← S ∪ {ui};
12          od;
13    fi;
```

Algorithmus 1: Bron-Kerbosch-Algorithmus

Ein Beispiel für eine Cliquensuche gibt Abbildung 11 wieder. Der Graph besitzt 4 Knoten und enthält 2 maximale Cliquen der Größe 3. Alle 4 Knoten dienen als Startpunkte für die Backtrackingsuche und stellen Verzweigungen auf der ersten Suchebene dar. Der Algorithmus vollzieht eine Tiefensuche und folgt zunächst immer dem ersten Zweig bis eine Gesamtlösung gefunden wurde. Die erste Lösung besteht aus dem Knotenset *C* = {1,2,3} und wird als beste bisherige Lösung abgespeichert. Anschließend erfolgt das Rücksetzten bis zur letzten Verzweigung. Die nächste Verzweigung führt zum Lösungset {1,3,4}, die zur ersten

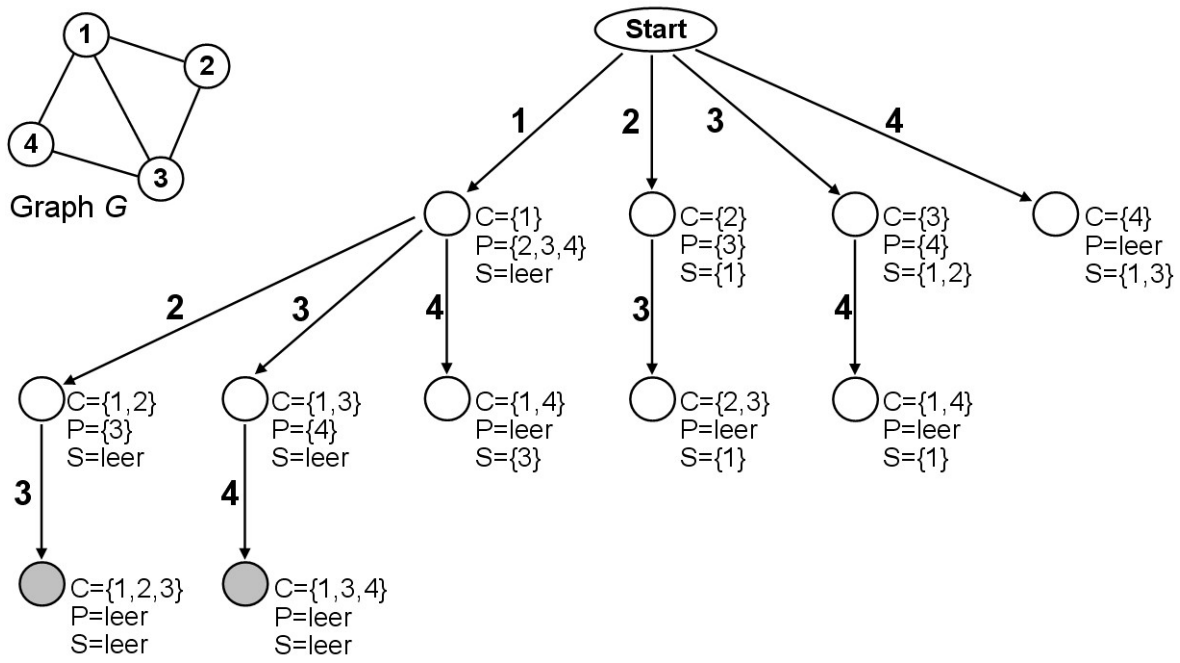


Abbildung 11: Backtracking Baum für die Suche nach den maximalen Cliquen von Graph G nach dem Bron-Kerbosch-Algorithmus.

Gesamtlösung hinzugefügt wird. Die nachfolgenden Verzweigungen werden nicht abgespeichert, da sie weniger Knoten beinhalten und Set *S* nicht leer ist. Wie sich anhand des Beispiels zeigt, verhindert Set S, dass Lösungen abgesucht werden, die bereits betrachtet wurden. Somit wird jede Clique nur einmal gefunden und die Laufzeit reduziert.

2.1.4 *c*-MCS-Algorithmus – eine Variante des Bron Kerbosch-Algorithmus

Der Bron-Kerbsoch-Algorithmus ist einer der schnellsten MCS Algorithmen (Brint *et al.*, 1987, Gerhards *et al.*, 1979) und läßt sich leicht modifiziern (Johnston, 1976). Eine Variante des Bron-Kerbosch-Algorithmus beschränkt sich auf die Suche nach zusammenhängenden maximalen gemeinsamen Subgraphen (Koch, 2001), die als *c*-MCS („*c*“ bedeutet „connected“, deutsch „zusammenhängend“) bezeichnet werden. Eine *c*-MCS wird durch eine so genannte *c*-Clique im Produktgraphen repräsentiert.

Eine besondere Bedeutung bei der Bildung der *c*-MCS kommt den *c*-Kanten im Produktgraphen zu. Die *c*-Kanten des Produktgraphen repräsentieren die Verbundenheit von Atomen in den beiden Ausgangsgraphen. Besteht eine Clique des Produktgraphen nur aus c-Kanten, so sind alle Knoten des korrespondierenden gemeinsamen Subgraphen miteinander verbunden. Die *d*-Kanten kommen hingegen zustande, wenn die jeweiligen Knoten in den Ausgangsgraphen nicht miteinander

verbunden sind. Sie repräsentieren Brüche in dem maximalen gemeinsamen Subgraphen. Enthält daher eine Clique nur *d*-Kanten, sind alle Knoten des gemeinsamen Subgraphen nicht miteinander verbunden. Die *d*-Kanten spielen bei der Bildung der *c*-MCS eine untergeordnete Rolle.

Dagegen kommt den so genannten *c*-Pfaden bei der Suche nach *c*-MCS eine zentrale Bedeutung zu. Ein *c*-Pfad ist eine aufeinander folgende Sequenz von *c*-Kanten, die nicht durch *d*-Kanten unterbrochen wird. 2 beliebige Knoten *u* und *v* einer *c*-Clique müssen über einen *c*-Pfad miteinander verknüpft sein.

Wie das Beispiel in Abbildung 12 zeigt, kann eine *c*-Clique neben *c*-Kanten auch *d*-Kanten enthalten. Die maximale *c*-Clique wird durch die Knoten {1,3,4,8} gebildet. Der zugehörige *c*-Pfad wird durch 3 *c*-Kanten gebildet. Die 3 anderen Kanten dieser Clique sind *d*-Kanten.

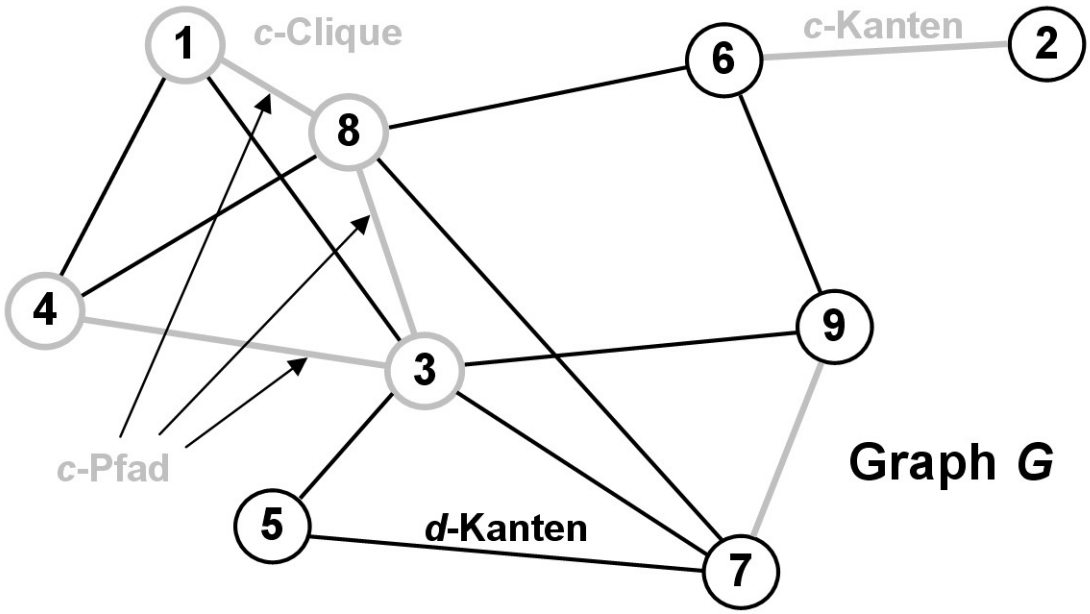


Abbildung 12: Zusammenhang zwischen *c*-Kanten, *c*-Pfaden und *c*-Cliquen eines Graphen *G* .

Der *c*-MCS-Algorithmus benutzt nur *c*-Kanten zur Erweiterung bestehender Teillösungen. Durch diese Eigenschaft folgt der Algorithmus den *c*-Kanten in Graph G und untersucht alle *c*-Pfade mit den zugehörigen *c*-Cliquen. Zu diesem Zweck unterscheidet die Bron-Kerbosch-Variante zwischen folgenden Sets:

Set *C*: beinhaltet Knoten, die bereits betrachtet und zum gegenwärtigen Cliquenset hinzugefügt wurden

Set *P*: Knoten, die im nächsten Suchschritt noch zu Set *C* hinzugefügt werden können. Sie sind mit dem Suchknoten u_i über eine ***c*-Kante** verbunden.

Set *D*: Knoten, die in einem der nächsten Suchschritte zu Set *C* hinzugefügt werden könnten. Sie sind mit dem Suchknoten u_i über eine ***d*-Kante** verbunden.

Set *S*: Knoten, deren Cliquen bereits betrachtet wurden und von denen bekannt ist, dass sie nicht zu Set *C* hinzugefügt werden können.

Gegenüber dem ursprünglichen Bron-Kerbosch-Algorithmus besitzt Set *P* nur Knoten, die über *c*-Kanten mit dem jeweiligen Suchknoten u_i verbunden sind. Andere Knoten, die hingegen über *d*-Kanten mit dem jeweiligen Suchknoten u_i verbunden sind, werden zunächst Set *D* zugeordnet. In Abhängigkeit davon, welcher Suchknoten u_i gewählt wird, können später Knoten aus Set *D* zu Set *P* übergehen, wenn sie mit dem Knoten u_i über eine *c*-Kante verbunden sind. Diese Zusammenhänge gehen auch aus dem Pseudocode des *c*-MCS-Algorithmus hervor. Zugrunde liegt der rekursive Aufruf der Funktion „Errechne_c_Cliquen“ :

```
Errechne_c_Cliquen(C,P,D,S)
> errechne alle c-Cliquen eines beliebigen Graphen G
C: Knoten der gegenwärtigen Clique
P: Knoten, über c-Kante mit u_i verbunden
D: Knoten, über d-Kante mit u_i verbunden
S: Knoten, die nicht zu Set C hinzugefügt werden können
N[u]: Set von Nachbarknoten von Knoten u in G
01    P is the Set {u_1,...,u_k}
02    if  P=∅ and S=∅
03          then report Clique;
04    else  for i ← 1 to k
05          do    P ← P\{u_i,};
06                P' ← P;
07                D' ← D;
08                S' ← S;
09                N ← {v ∈ V | {u_i,v} ∈ E}
10                for all v ∈ D'
11                      do if  sind v und u_i benachbart über c-Kante
12                            then P' ← P' ∪ {v};
13                            then D' ← D'  \ {v};
14                      fi;
15                      od;
16                Errechne_c_Cliquen(C ∪ {u_i}, P'∩N, D'∩N, S'∩N);
17                S ← S ∪ {u_i};
18          od;
19    fi;
```

Algorithmus 2: *c*-MCS-Algorithmus

Die Funktion „Errechne_c_Cliquen“ wird zu Beginn durch die Initialisierungsfunktion „Init_Algorithmus_2“ aufgerufen (Algorithmus 3). Diese Funktion betrachtet alle Knoten von Graph *G* als mögliche Startknoten für die *c*-MCS-Suche. Je nachdem welcher Startknoten *u* aus dem Gesamtset *V* gewählt wird, werden verschiedene Startsets für *P*, *D* und *S* gebildet. Die Funktion von Set *S* in den Algorithmen 1 und 2 wird in „Init_Algorithmus_2“ teilweise von Set *T* übernommen. Knoten, die bereits betrachtet wurden, werden in Set T aufgenommen (Zeile 18).
In Zeile 06 werden alle Nachbarknoten von Knoten *u* ermittelt und in der folgenden *for*-Schleife (Zeile 07 bis 15) den verschiedenen Sets zugeordnet. Nachbarknoten, die über eine *c*-Kante mit *u* verbunden sind und bereits betrachtet wurden, werden Set *S* zugeordnet (Zeile 09 und 10). Nur Nachbarknoten, die noch nicht betrachtet wurden und über eine *c*-Kante mit Knoten *u* verbunden sind, werden Set *P* zugeordnet. Set *D* erhält die Nachbarknoten, die eine *d*-Kante mit dem Startknoten *u* teilen (Zeile 13 und 14).

Init_Algorithmus_2 ()

Funktion: Initialisierung von Algorithmus 2
C: Enthält Startknoten für die Cliquensuche
P: Knoten, die über *c*-Kante mit u_i verbunden sind
D: Knoten, die über *d*-Kante mit u_i verbunden sind
T: Knoten, die nicht zu Set *C* hinzugefügt werden können
N[*u*]: Set von Nachbarknoten von Knoten *u* in *G*

```
01    T ← ∅;
02    for all u ∈ V
03          do
04                P ← ∅;
05                S ← ∅;
06                N← {v ∈ V | {u,v} ∈ E};
07                for each v ∈ N
08                      do if u and v sind benachbart über c-Kante
09                            if  v ∈ T
10                                  then  S ← S ∪ {v};
11                            else  P ← P ∪ {v};
12                            fi;
13                      else if u and v sind benachbart über d-Kante
14                            then  D ← D ∪ {v};
15                      fi od;
16          od;
17          Errechne_c_Cliquen({u},P,D,S);
18          T ← T ∪ {u};
```

Algorithmus 3: Initialisierung von Algorithmus 2

Der *c*-MCS-Algorithmus ist durch die Einschränkung der Suche auf zusammenhängende maximale gemeinsame Subgraphen wesentlich schneller als der ursprüngliche Bron-Kerbosch-Algorithmus. Der Geschwindigkeitsvorteil macht sich umso mehr bemerkbar, je größer die beiden Eingangsgraphen bzw. die Moleküle werden. Wenn die Eingangsgraphen groß sind, steigt die Wahrscheinlichkeit an, dass die Knoten im Produktgraphen durch *d*-Kanten miteinander verbunden sind. Wie in Tabelle 2, aufgeführt besitzt der Produktgraph bei dem Vergleich von Propanol mit Isopropanol nur 6 *d*-Kanten und 9 *c*-Kanten. Damit gibt es mehr *c*- als *d*-Kanten. Dies begründet sich mit der großen Ähnlichkeit und der geringen Größe der beiden Moleküle. ATP besitzt hingegen bereits 31 Atome und AMP 23 Atome ohne Berücksichtigung der Wasserstoffatome. Obwohl ATP und AMP sehr ähnliche Moleküle sind, besitzt der Produktgraph bereits 18974 *d*-Kanten, aber nur 245 *c*-Kanten. FAD bzw. $FADH_2$ bestehen aus 53 Rückgratatomen. Der zugehörige Produktgraph besitzt 471127 *d*-Kanten und 1346 *c*-Kanten. Diese Zahlen machen den Geschwindigkeitsvorteil deutlich, der sich aus der Suche nach den maximalen Cliquen mit den längsten *c*-Pfaden ergibt.

Vergleichsmoleküle		*d*-Kanten (Anzahl)	*c*-Kanten (Anzahl)
Propanol	Isopropanol	6	9
ATP	AMP	18974	245
$FADH_2$	FAD	471127	1346

Tabelle 2: Beziehung zwischen der Anzahl der *c*- und *d*-Kanten des Produktgraphen und der Größe der beiden Vergleichsmoleküle.

Obwohl der *c*-MCS-Algorithmus nur zusammenhängende maximale gemeinsame Subgraphen findet, erwies sich diese Eigenschaft für die Generierung einer Atom-Atom-Zuordnung von Reaktionen als besonders vorteilhaft. Wie in Abschnitt 2.2.2 noch genauer erläutert wird, war für die Atomzuordnung komplexerer Reaktionen eine Bewertung der MCS nach ihrer Größe erforderlich. Allerdings führt diese Bewertungsmethode nur zu signifikanten Ergebnissen, wenn zusammenhängende MCS betrachtet werden.

Nachteile des *c*-MCS-Algorithmus:

Auch wenn der *c*-MCS-Algorithmus mit einer deutlich günstigeren Laufzeit einhergeht, bleibt die Laufzeit exponentiell. Somit ist nicht der Vergleich beliebig großer Moleküle möglich. Die Laufzeit ist abhängig von der Größe der zu vergleichenden Moleküle, ihrer Elementzusammensetzung, der Molekülstruktur und anderer Faktoren. Ein effektiver Vergleich großer Metabolite ist mit dem *c*-MCS-Algorithmus daher noch nicht möglich.

Ein weiterer Nachteil des Algorithmus zeigt sich bei der Rekonstruktion komplexerer Reaktionen, bei denen aromatische Ringe oder andere Ringssysteme gespalten oder gebildet werden. Ein einfacheres Beispiel hierfür ist die Reaktion der Catechol:oxygen 1,2-oxidoreductase, in der ein aromatischer Ring gespalten wird (Abbildung 13 A). Unter der Beteiligung eines Wassermoleküls wird Catechol zu Muconat gespalten. Dabei gehen alle 8 Atome von Catechol in das Muconat über. Der *c*-MCS-Algorithmus generiert als Variante des Bron-Kerbosch-Algorithmus allerdings 12 MCS, wobei jeweils nur 6 Atome des Catechols dem Muconat zugeordnet werden (Abbildung 13 B).

Es stellte sich heraus, dass die Ursache hierfür bereits bei der Generierung des Knoten-Produktgraphen H_v entsteht. Dabei werden Einschränkungen vorgenommen, die teilweise zum Fehlen entsprechender Kanten im Produktgraphen führen. Der Produktgraph stellt aber die Grundlage für alle Algorithmen dar, die das MCS-Problem mit Hilfe des Cliquen-Problems lösen.

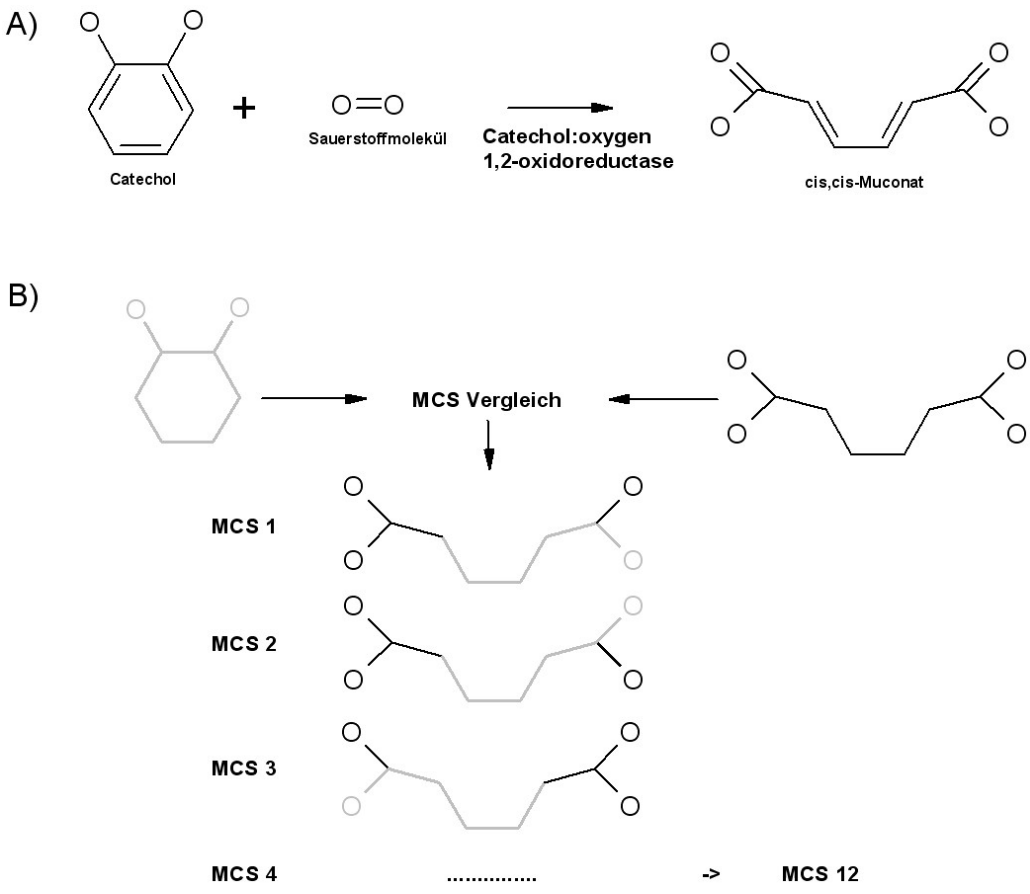


Abbildung 13: A) Reaktion der Catechol:oxygen 1,2-oxidoreductase. B) Eine MCS-Suche mit Hilfe des *c*-MCS-Algorithmus führt zu 12 MCS, wobei jeweils nur 6 Atome des Catechols dem Muconat zugeordnet werden können.

2.1.5 McGregor-Algorithmus

Im Zuge dieser Arbeit wurden noch weitere MCS Algorithmen getestet. Ein sehr flexibler MCS-Algorithmus, der für den Vergleich von Molekülen entwickelt wurde, geht zurück auf (McGregor, 1982). Bei diesem Algorithmus werden mittels eines Backtracking-Verfahrens systematisch alle möglichen Zuordnungen der Knoten des einen Graphen auf den anderen Graphen ausprobiert. Die Bindungen werden als Knoten betrachtet und die Bindungen des einen Moleküls werden gegen die Bindungen des anderen Moleküls in einer zweidimensionalen Matrix aufgetragen. Abbildung 14 zeigt den Aufbau der zweidimensionalen Startmatrix für den Vergleich der Moleküle Propanol und Isopropanol. Beide Moleküle verfügen über eine C-O und zwei C-C-Bindungen. Da die C-C-Bindungen von Propanol den C-C-Bindungen von Isopropanol zugeordnet werden können, werden die entsprechenden Positionen in der Startmatrix mit einem Wert von „1“ markiert. Auch die Position, welche die Übereinstimmung der C-O Bindung von Propanol und Isopropanol kennzeichnet, erhält den Eintrag einer „1“. Ein Eintrag von „1“ in der Startmatrix repräsentiert also eine mögliche Zuordnung und wird im folgenden Text auch als „Zuordnung“ bezeichnet. Die restlichen Positionen erhalten einen Wert von „0“ als Zeichen dafür, dass hier keine Zuordnung möglich ist.

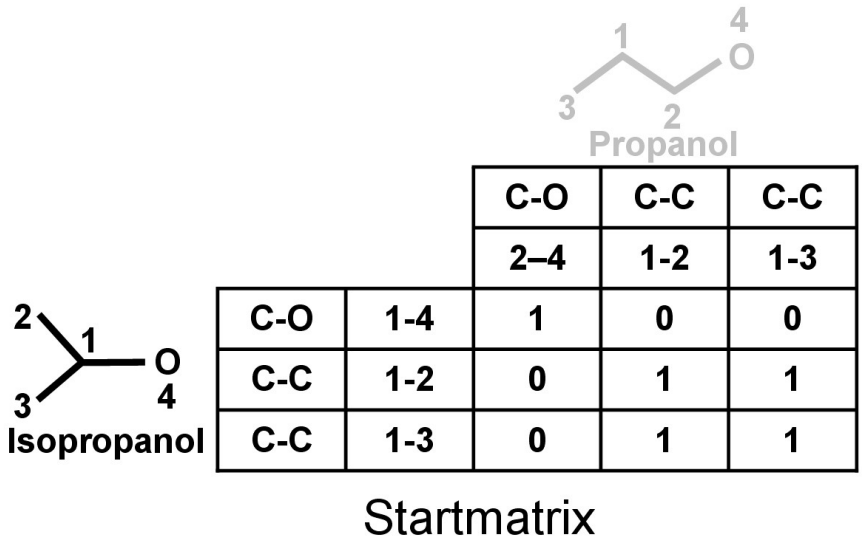


Abbildung 14: Startmatrix des McGregor-Algorithmus für die MCS-Suche von Propanol mit Isopropanol.

Anschließend erfolgt eine Backtrackingsuche, die auf dem rekursiven Aufruf einer Funktion basiert, ebenso wie bei dem in Abschnitt 2.1.3 beschriebenen Bron-Kerbosch-Algorithmus. An jedem

Knotenpunkt innerhalb des Backtrackingsuchbaumes legt sich der Algorithmus auf mögliche Zuordnungen in der Matrix fest. Jedes mal, wenn eine Zuordnung gewählt wird, können andere Zuordnungen ausgeschlossen werden. Die entsprechenden Zuordnungen in der Matrix werden dann gelöscht und es wird zu der nächsten Zuordnungen übergegangen, die übrig geblieben sind und noch nicht betrachtet wurde.

Gemäß des McGregor-Algorithmus, stellt jede in der Matrix verbliebene und noch nicht betrachtete Zuordnung bei jedem Knoten innerhalb des Suchbaumes eine weitere Verzweigung dar. Dies führt zu einer explosionsartigen Ausweitung des Suchbaumes. Suchpfade werden in unterschiedlicher Reihenfolge der Knoten in hoher Anzahl betrachtet und Lösungen vielfach gefunden. Durch diese rechenaufwendige Operation ist nur der Vergleich von kleinen Molekülen möglich. Aus diesem Grund wurde ein „divide and conquer"-System (deutsch „teile und herrsche") ausgearbeitet. Dieses ermöglicht, dass jede Matrixkonstellation nur einmal betrachtet und jede Lösung nur einmal gefunden wird.

Das „divide and conquer"-System beruht auf dem Umstand, dass jede Zuordnung in der Matrix 2 Zustände annehmen kann. Zum einen kann eine Zuordnung als Teillösung angesehen werden und den Wert „1" behalten. Die andere Möglichkeit ist, dass eine Zuordnung den Wert „0" erhält und somit nicht als Teillösung betrachtet wird. In Hinblick auf diese beiden Zustände, ergibt sich die Möglichkeit, die Anzahl der Verzweigungen auf 2 zu reduzieren, wenn eine Zuordnung betrachtet wird. In der ersten Verzweigung wird die rekursive Funktion erneut aufgerufen unter der Voraussetzung, dass die Zuordnung Teil einer Gesamtlösung ist. Dabei werden alle Zuordnungen, die sich nicht mit der Gewählten vereinbaren lassen, gestrichen und es wird zur nächsten verbliebenen Zuordnung übergegangen. In der zweiten Verzweigung wird die Zuordnung nicht als Teil einer Gesamtlösung angesehen. Daher wird die Matrixposition auf „0" gesetzt und es wird zur nächsten Zuordnung übergegangen.

Abbildung 15 zeigt die 6 Lösungsmatrizen, die bei dem Vergleich von Propanol mit Isopropanol entstehen. Jede Lösungsmatrix repräsentiert eine MCS, deren Struktur neben der Lösungsmatrix angegeben wird. Abbildung 16 veranschaulicht die Funktionsweise des „divide and conquer"-Systems. Nicht jede Zuordnung innerhalb der Startmatrix dient als Anfangspunkt für eine Backtrackingsuche. Das System ermöglicht es nur die erste Zuordnung als Ausgangsposition zu verwenden, da auch die Möglichkeit betrachtet wird, dass diese Zuordnung nicht Teil einer Gesamtlösung sein muss. Hierdurch lassen sich weiterhin alle möglichen Matrixkonstellationen erzeugen. Dennoch werden alle möglichen

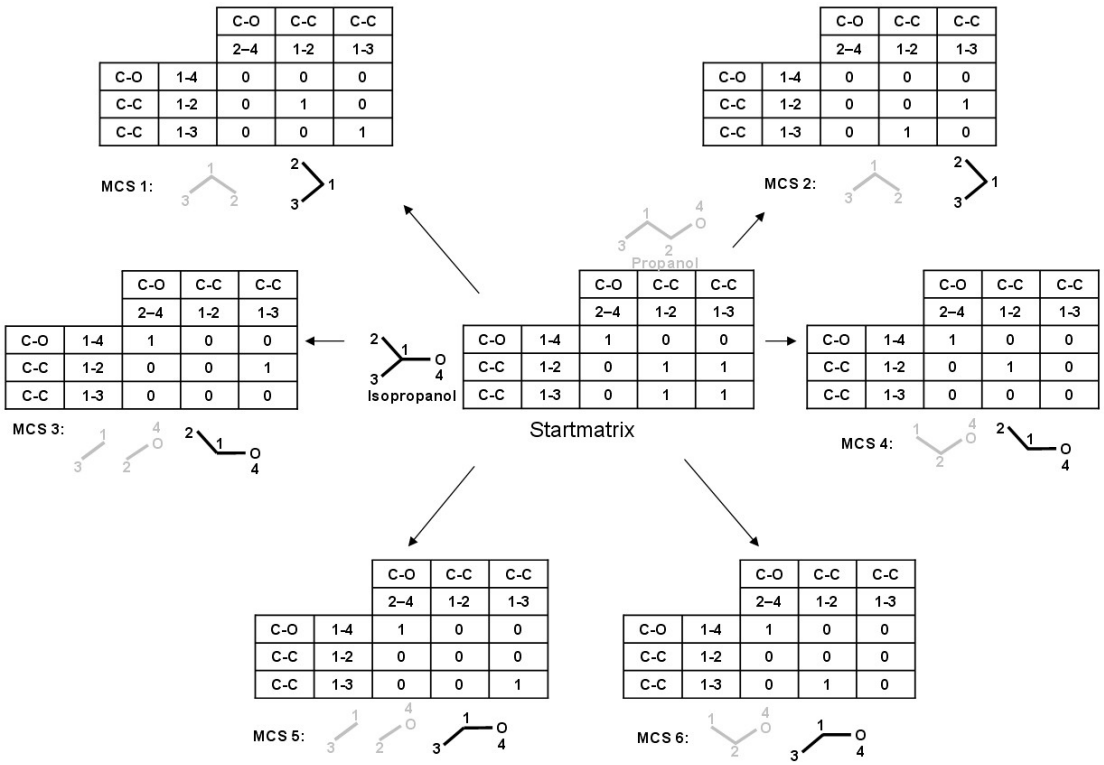


Abbildung 15: Startmatrix und Lösungsmatrizen des McGregor-Algorithmus für den Vergleich von Isopropanol und Propanol.

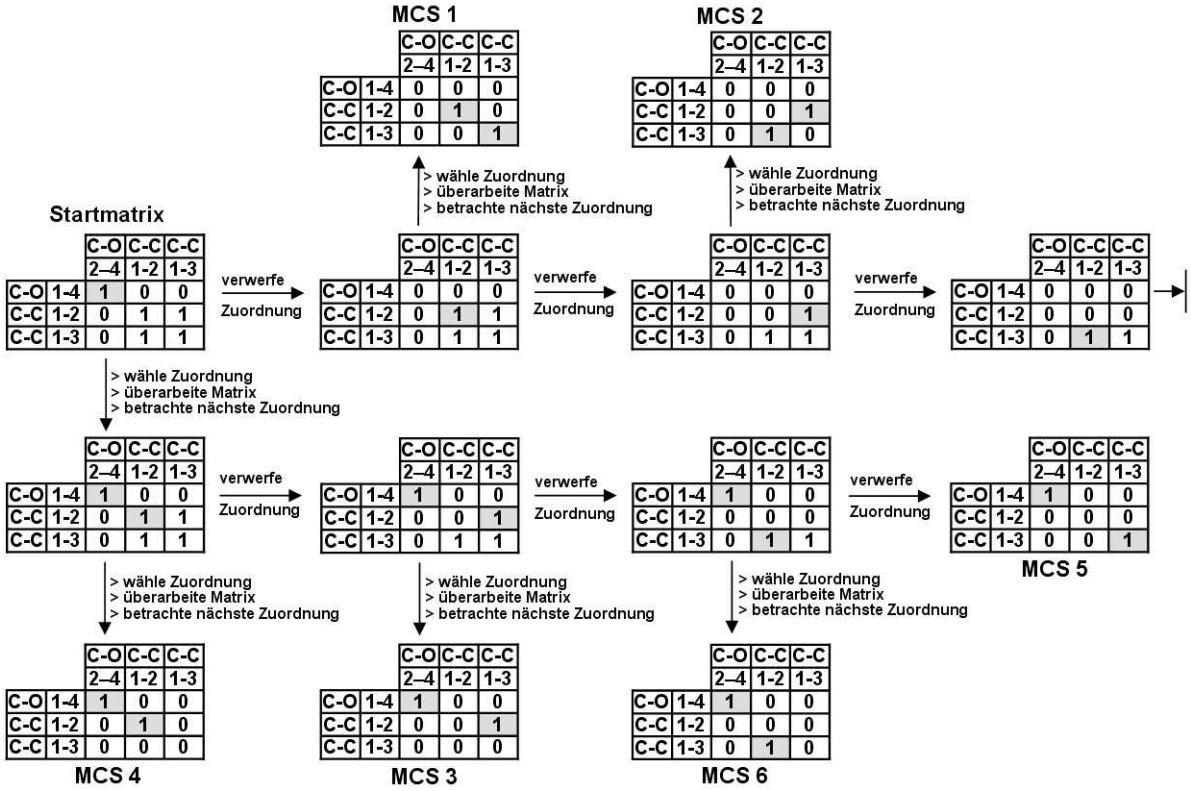


Abbildung 16: Suchschema des „divide and conquer“-Systems für den Vergleich von Isopropanol und Propanol.

Lösungswege betrachtet und der Algorithmus findet alle MCS des Beispiels nach wenigen rekursiven Schritten.

Im Gegensatz zu den Cliquenalgorithmen, die auf dem Knoten-Produktgraphen H_v basieren, findet der McGregor-Algorithmus 6 MCS. Der Knoten-Produktgraphen H_v enthält hingegen nur 4 maximale Cliquen. Algorithmen, die auf einer Zuordnung der Bindungen beruhen, erlauben ein höheres Maß an Flexibilität (Abschnitt 4.1.2.4). Der Vergleich von Catechol und Muconat mit Hilfe des McGregor-Algorithmus führt zu 8 MCS, wobei jeweils alle Atome des Catechol vollständig den Atomen von Muconat zugeordnet werden (Abbildung 17).

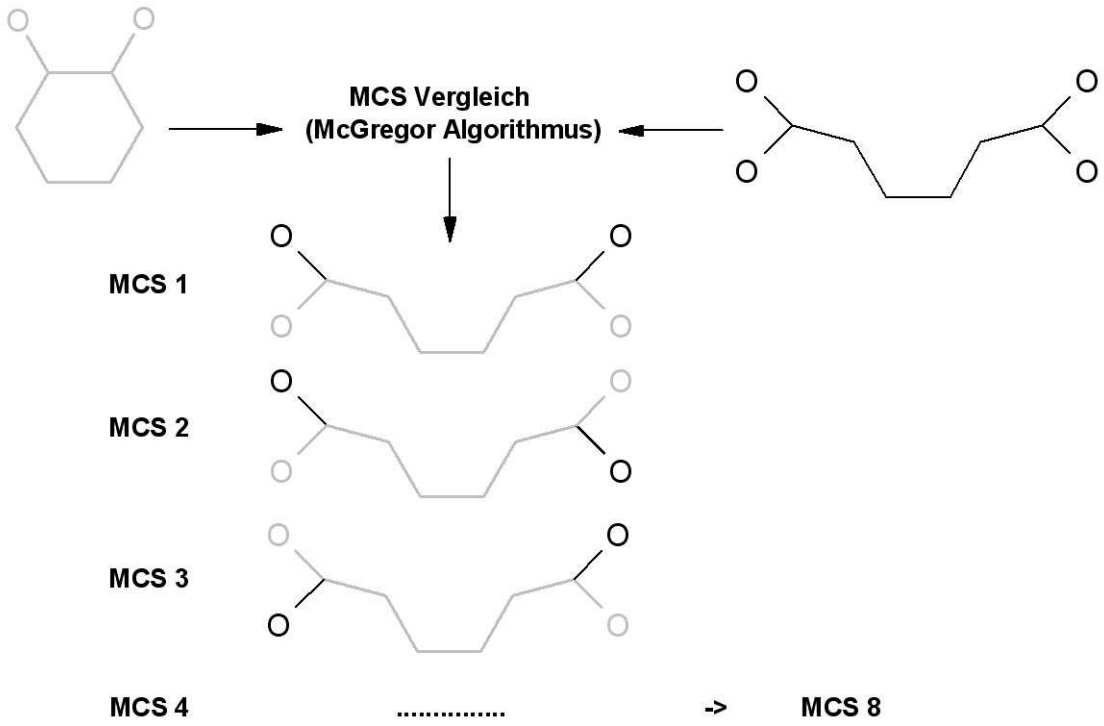


Abbildung 17: Eine MCS-Suche mit Hilfe des McGregor-Algorithmus führt zu einer vollständigen Zuordnung aller Catechol-Atome auf die Atome von Muconat.

Nachteile des McGregor-Algorithmus:

Der McGregor-Algorithmus wird zur Zuordnung von Bindungen benutzt. Einzelne Atome können daher durch den Algorithmus nicht zugeordnet werden. Die Zuordnung einzelner Atome benötigt daher Zusatzfunktionen. Auch das Auslesen der Atomzuordnungen aus den Bindungszuordnungen erfordert einen Zusatzalgorithmus.

Die Laufzeit ist exponentiell, wie bei allen anderen MCS Algorithmen. Ein Vergleich von Metaboliten mittlerer Größe ist somit bereits problematisch.

2.1.6 Kombination des *c*-MCS- und McGregor-Algorithmus

Nachdem zunächst Versuche unternommen wurden, die Algorithmen einzeln zu optimieren, ergab sich die Idee, die Schnelligkeit des *c*-MCS-Algorithmus mit der Flexibilität des McGregor-Algorithmus zu kombinieren.

Zu Beginn wird dem schnelleren *c*-MCS-Algorithmus Priorität eingeräumt. Zwei Moleküle werden zunächst mit Hilfe des *c*-MCS-Algorithmus verglichen. In vielen Fällen führt der Vergleich bereits zu *c*-MCS, die sich nicht mehr erweitern lassen. Dies ist besonders offensichtlich, wenn sich bereits alle Atome des einen Moleküls in das andere Molekül übertragen lassen. Bleiben jedoch in beiden Molekülen Atome und Bindungen gleichen Typs zurück, werden diese an den McGregor-Algorithmus weitergegeben, der die Suche fortsetzt. Damit weiterhin nach zusammenhängenden maximalen gemeinsamen Substrukturen, *c*-MCS, gesucht wird, werden nicht alle verbliebenen Bindungen auf einmal an den McGregor-Algorithmus weitergegeben. Vielmehr werden in einem iterativen Verfahren immer nur die Bindungen an den McGregor-Algorithmus weitergegeben, die unmittelbar an die bisher gefundene gemeinsame Substruktur angrenzen. Diese Vorgehensweise wird in Abbildung 18 am Beispiel des Vergleichs von Catechol und Muconat näher erläutert. Zunächst wird durch den ursprünglichen *c*-MCS-Algorithmus eine gemeinsame Substruktur ermittelt. Ein Vergleich der verbliebenen Bindungen zeigt allerdings, dass noch Bindungen gleichen Typs in beiden Molekülen vorhanden sind. Die Bindungen, die an die gemeinsame Substruktur angrenzen, werden ermittelt und dem McGregor-Algorithmus übergeben.

An diesen Bindungen sind Atome im Grenzbereich der bisherigen gemeinsamen Substruktur beteiligt. Die Grenzatome müssen die bisherige Atomzuordnung beibehalten und erhalten daher ein spezifisches Label anstelle des Elementsymbols. Zum Beispiel ist Atom 1 von Catechol dem Atom 8 von Muconat zugeordnet. Diese Atome erhalten das Label „X“. Ebenso ist Atom 5 von Catechol dem Atom 2 von Muconat zugeordnet. Entsprechend erhalten diese Atome das Label „Q“.

Im ersten Suchschritt erweitert der McGregor-Algorithmus die gemeinsame Substruktur um eine Q-C Bindung. Anschließend werden die Grenzbereiche der gemeinsamen Substruktur neu definiert und Atom 2 von Catechol und Atom 1 von Muconat erhalten das Label „Q“. Es ergeben sich nun 2 Möglichkeiten, die gemeinsame Substruktur um eine Q-O Bindung auszubauen. Danach lässt sich die gemeinsame Substruktur nicht mehr erweitern und es werden 2 *c*-MCS ausgegeben. Ebenso baut der McGregor-Algorithmus die anderen gemeinsamen Substrukturen aus, die als Grundlage von dem ursprünglichen *c*-MCS-Algorithmus errechnet werden. Am Ende werden 8 *c*-MCS ermittelt.

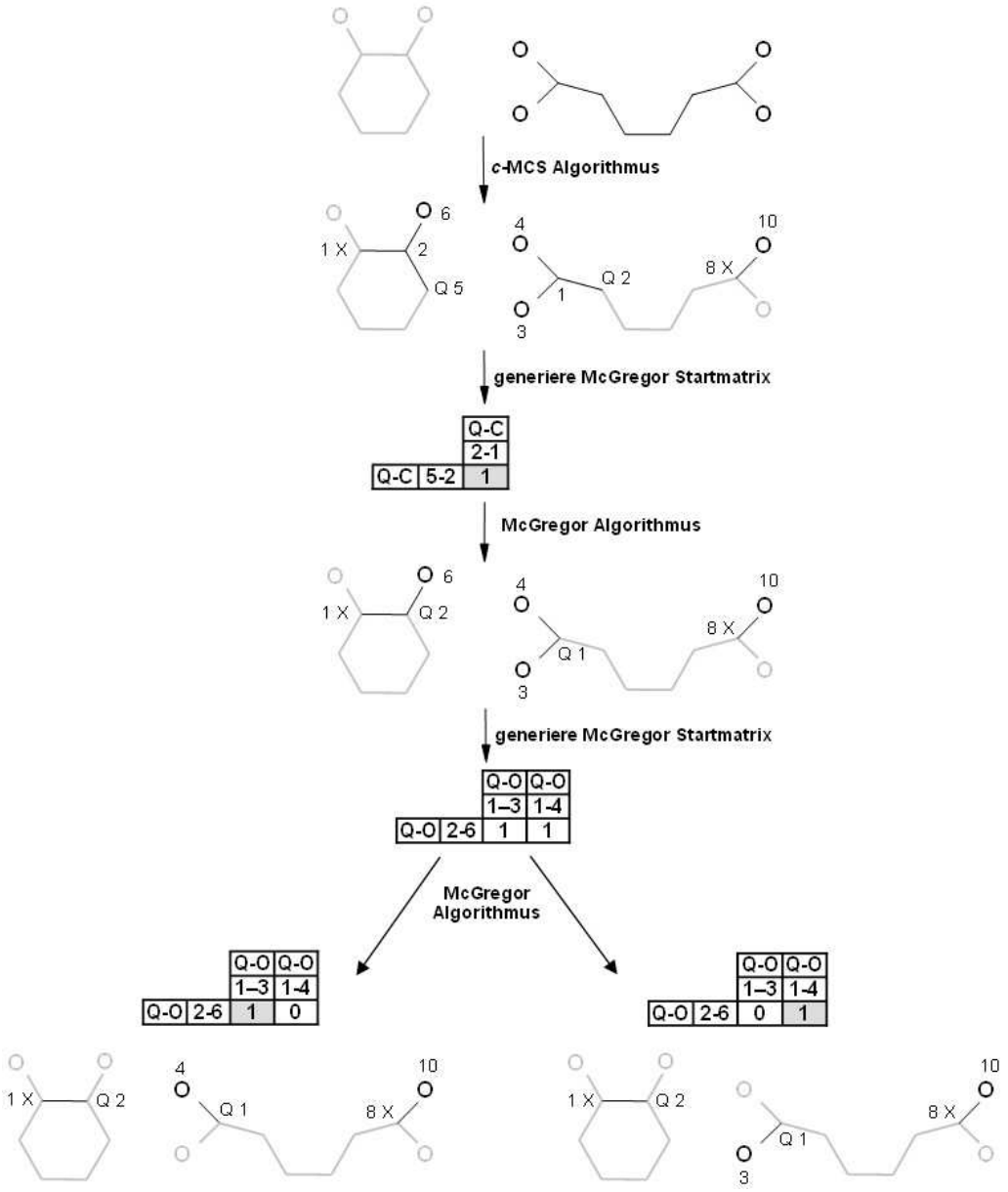


Abbildung 18: Funktionsweise des flexibleren *c*-MCS-Algorithmus, der aus der Verbindung des ursprünglichen *c*-MCS-Algorithmus und des McGregor-Algorithmus hervorgeht.

Nicht in allen Fällen ist die Erweiterung der gemeinsamen Substrukturen so übersichtlich, wie in dem Beispiel von Abbildung 18. Häufig gibt es zahlreiche Möglichkeiten die zugrunde liegende Substruktur auszubauen. Aus diesem Grund werden durch ein Backtrackingverfahren alle Möglichkeiten ausgetestet, die Substruktur durch den McGregor-Algorithmus zu erweitern. Mit dem ursprünglichen *c*-MCS-Algorithmus und dem McGregor-Algorithmus beinhaltet der neue *c*-MCS-Algorithmus somit 3 verschachtelte Backtrackingverfahren. Das Ergebnis ist andererseits ein sehr flexibler und schneller *c*-MCS-Algorithmus.

Verbesserung der Laufzeit:

Die Kombination des ursprünglichen *c*-MCS-Algorithmus mit dem McGregor-Algorithmus wurde auch dazu benutzt, die Laufzeit für den Vergleich von großen Molekülen zu verbessern. Der Ansatz beruht auf der Annahme, dass es in einem ersten Schritt ausreichen sollte, gemeinsame Substrukturen zu errechnen, da diese anschließend mit Hilfe des McGregor-Algorithmus bis zur *c*-MCS erweitert werden können. Voraussetzung hierfür ist eine ausreichende Größe der gemeinsamen Substrukturen, die durch den ursprünglichen *c*-MCS-Algorithmus als Startstrukturen vorgegeben werden. Hierdurch wird die Wahrscheinlichkeit erhöht, dass die Startstruktur im Bereich der tatsächlichen MCS der Moleküle liegt. In diesem Fall lassen sich größere Einschränkungen in den Zuordnungsregeln des vorgeschalteten *c*-MCS-Algorithmus vornehmen.

Durch striktere Zuordnungsregeln kann der Knoten-Produkt-Graph H_v erheblich verkleinert werden. In dem ursprünglichen *c*-MCS-Algorithmus wurde nur der Elementtyp eines Atoms berücksichtigt (Abschnitt 2.1.3). Jedes Kohlenstoffatom des einen Moleküls bildet beispielsweise mit jedem Kohlenstoffatom des anderen Moleküls einen Knoten im Produktgraphen. Bei den neuen Zuordnungsregeln wird bei größeren Molekülen auch die Umgebung der Atome berücksichtigt. Jedes Atom erhält zunächst ein Label. Dieses besteht aus einer Zahlenfolge von 7 Zahlen. Jedem Elementtyp wird eine bestimmte Zahl zugeordnet.

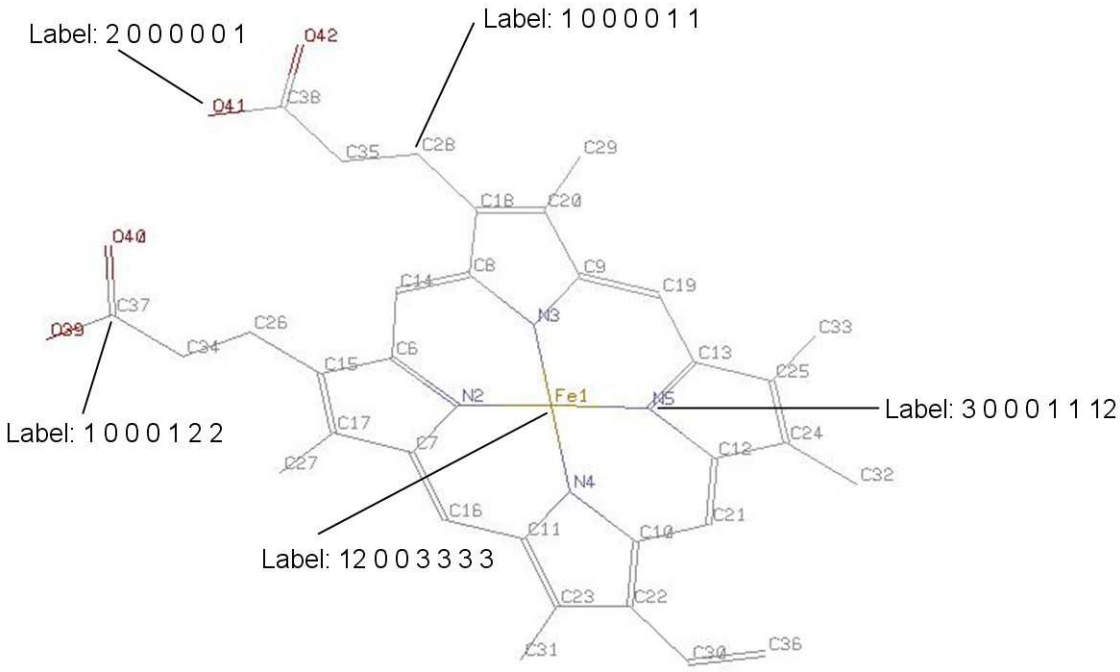


Abbildung 19: Die Atome des Ferrocytochroms erhalten Label, die Aufschluss über die atomare Umgebung geben.

In Abhängigkeit davon, von welchen Elementen ein Atom umgeben ist, setzt sich die Zahlenfolge anders zusammen. Nur Atome mit identischem Label werden im Produktgraphen H_v zu neuen Knoten zusammengefasst. Die Größe des Produktgraphen H_v nimmt hierdurch deutlich ab. Die Verbesserung in der Laufzeit ist so erheblich, dass auch der Vergleich von großen Metaboliten in akzeptabler Zeit erfolgen kann. Bei großen Metaboliten beträgt die *c*-MCS-Suche teilweise noch mehr als eine Minute mit einem Pentium-Rechner (2,4 GHz). Beispielsweise benötigt der Vergleich von GM1 und GM2 80 Sekunden. Ohne Berücksichtigung der Wasserstoffatome besitzt GM1 91 Atome und GM2 80 Atome. Die Laufzeit hängt aber nicht nur von der Molekülgröße, sondern auch von der Elementzusammensetzung und der Molekülstruktur ab. So benötigt der Vergleich von Vitamin B12 und Cob(II)alamin nur 23 Sekunden, obwohl beide Moleküle ohne Berücksichtigung der Wasserstoffatome eine Größe von 91 Atomen besitzen. Für mittelgroße Moleküle, wie beispielsweise FAD und $FADH_2$ mit 53 Atomen, beträgt die Laufzeit bereits deutlich weniger als eine Sekunde. Diese und kleinere Moleküle machen die Mehrheit der Metabolite aus.

Bei dem Vergleich von kleinen Molekülen ist die Berücksichtigung der atomaren Umgebung nicht sinnvoll. Zum einen ist die Laufzeit auch ohne Berücksichtigung der Umgebung ohnehin gering. Andererseits ändert sich im Verlauf von biochemischen Reaktionen die atomare Umgebung innerhalb kleiner Moleküle häufig sehr. Die Wahrscheinlichkeit, falsche Zuordnungen zu treffen, ist bei kleinen Molekülen daher zu hoch. Daher verzweigt sich das Programm in Abhängigkeit von der Größe der betrachteten Moleküle. Bei dem Vergleich von kleinen Molekülen erfolgt bei dem *c*-MCS-Algorithmus, der dem McGregor-Algorithmus vorgeschaltet ist, keine Betrachtung der atomaren Umgebung. Überschreiten aber beide Eingangsmoleküle eine Größe von 40 Atomen, so werden bei dem vorgeschalteten ursprünglichen *c*-MCS die atomaren Umgebungen berücksichtigt.

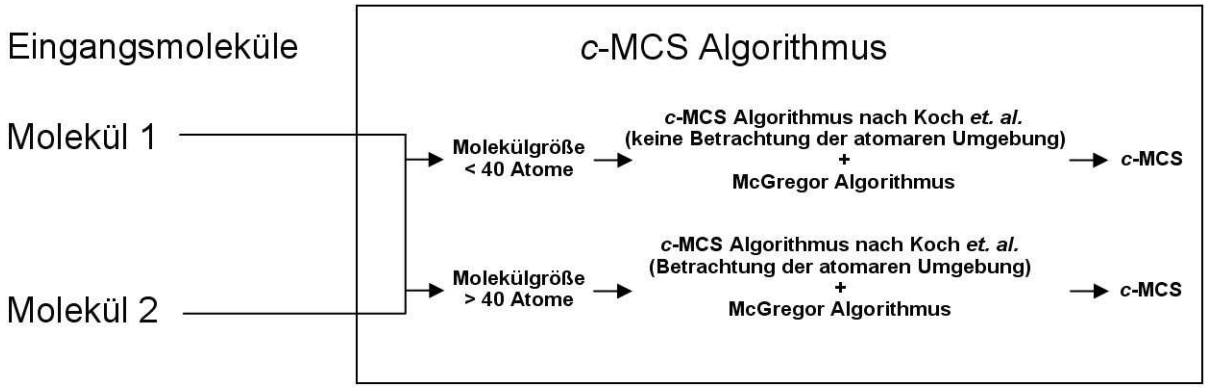


Abbildung 20: Der neue *c*-MCS-Algorithmus verzweigt sich in Abhängigkeit von der Molekülgröße.

2.2 Atom-Zuordnung biochemischer Reaktionen

Der unter Abschnitt 2.1 beschriebene *c*-MCS-Algorithmus dient als Grundlage, um Atomzuordnungen für biochemische Reaktionen zu generieren. Zu diesem Zweck wird mit Hilfe des *c*-MCS-Algorithmus jedes Eduktmolekül mit jedem Produktmolekül verglichen. Jeder einzelne Molekülvergleich führt zumeist zu einem Set von *c*-MCS. Nur in wenigen Fällen besitzen 2 Moleküle eine eindeutige *c*-MCS. Bei einer Reaktion mit mehreren Edukt- und Produktmolekülen entstehen auf diese Weise mehrere Sets von *c*-MCS. Um aus diesen Sets eine Atomzuordnung für die Gesamtreaktion zu generieren, müssen schließlich die *c*-MCS aus den verschiedenen Sets in allen möglichen Kombinationen ausgetestet werden. Ziel ist es dabei eine Kombination von *c*-MCS zu finden, bei der alle Edukt- und Produktatome einander zugeordnet werden können.
Je mehr Edukt- und Produktmoleküle an der Reaktion beteiligt sind, desto mehr nimmt auch die Anzahl an *c*-MCS Sets zu und die Kombinationsmöglichkeiten vervielfachen sich. Um auch in komplexen Fällen mit überlappenden *c*-MCS eine Zuordnung der Edukt- und Produktatome gewährleisten zu können, wurde ein Rankingsystem zur Bewertung von *c*-MCS für das Modul entwickelt. Auch heuristische Methoden helfen, die Anzahl möglicher Kombinationen zu verringern. So werden lokale Symmetrien innerhalb der Moleküle erkannt und redundante *c*-MCS aus den Sets entfernt. Kofaktoren werden separat betrachtet, um ihre Atome unabhängig von den übrigen Molekülen zuzuordnen. Um falsche Zuordnungen zu vermeiden, war es weiterhin erforderlich, einen Algorithmus zur Erkennung von Aromaten und anderen Ringsystemen zu entwickeln. Schließlich wurde auch eine Variante des Bron-Kerbosch-Algorithmus hinzugefügt, die es ermöglicht die Zuordnung von Molekülen und einzelnen Atomen zu trennen. Die Funktionsweise der Atomzuordnung und die einzelnen Optimierungsverfahren werden in den folgenden Abschnitten genauer erläutert.

2.2.1 *c*-MCS-Kombination

Der neue *c*-MCS-Algorithmus stellt den Kern des Atom-Atom-Zuordnungsprogramms dar. Er wird dazu verwendet, jedes Eduktmolekül mit jedem Produktmolekül zu vergleichen. Ein einfacheres Beispiel für die Generierung einer Atomzuordnung stellt die Reaktion der Pyruvatdecarboxylase dar (Abbildung 21). Dieses Enzym ist für die alkoholische Gärung verantwortlich und decarboxyliert Pyruvat unter Entstehung von Kohlendioxid und Acetaldehyd. Die Reaktion kann stellvertretend für Reaktionen des Typs A $\leftrightarrow$ B + C angesehen werden. Pyruvat wird als Eduktmolekül mit den Produktmolekülen Kohlendioxid und Acetaldehyd verglichen. Durch die beiden Molekülvergleiche

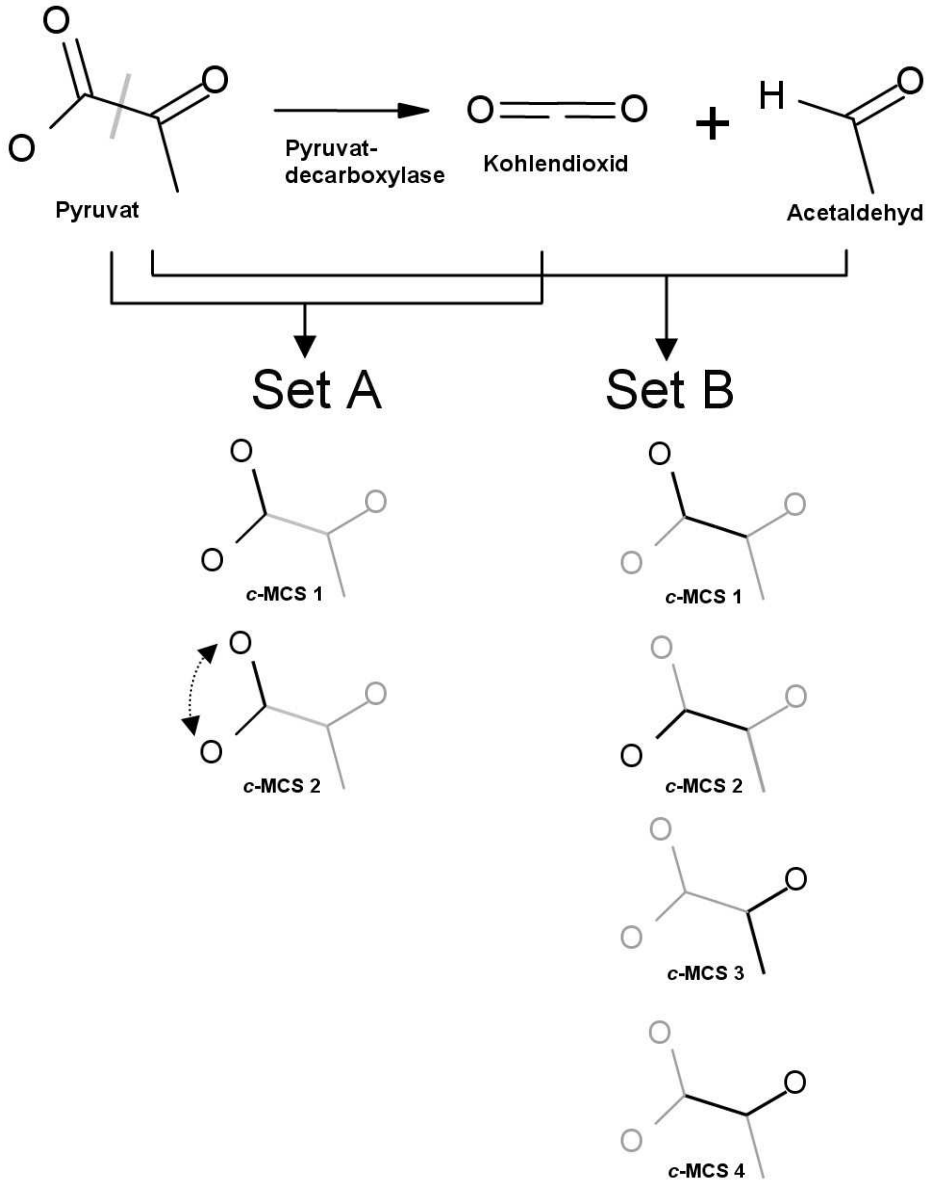


Abbildung 21: Edukt-Produkt *c*-MCS-Vergleichsschema anhand des Beispiels der Pyruvatdecarboxylase. Es entstehen die beiden Sets A und B mit jeweils mehreren *c*-MCS.

entstehen 2 Sets von *c*-MCS. Der Vergleich von Pyruvat mit Kohlendioxid führt zu 2 *c*-MCS (Set A), welche sich nur in der Orientierung unterscheiden, aber die gleichen Molekülbereiche abdecken. Set B geht aus dem Vergleich von Pyruvat mit Acetaldehyd hervor und beinhaltet 4 *c*-MCS. Um nun eine atomare Abbildung der Eduktatome auf die Produktatome generieren zu können, stellt sich die Frage, welche *c*-MCS aus Set A und Set B kombiniert werden müssen, damit eine vollständige Atomzuordnung entsteht. Zu diesem Zweck werden alle *c*-MCS-Kombinationen aus Set A und Set B generiert. Anschließend werden die Kombinationen selektiert, die zu einer maximalen Anzahl von Atomzuordnungen führen. Es entstehen 8 *c*-MCS-Kombinationen. Nur 2 Kombinationen führen zu

einer vollständigen Atomzuordnung, während die restlichen Kombinationen Atomüberlagerungen enthalten. Reaktionen, in denen ein Metabolit in zwei Molekülfragmente zerfällt, lassen sich einfach nachvollziehen. In diesem Fall lassen sich immer Kombinationen finden, wo sich die beiden Fragmente ohne Überlagerung in das Ausgangsmolekül übertragen lassen.

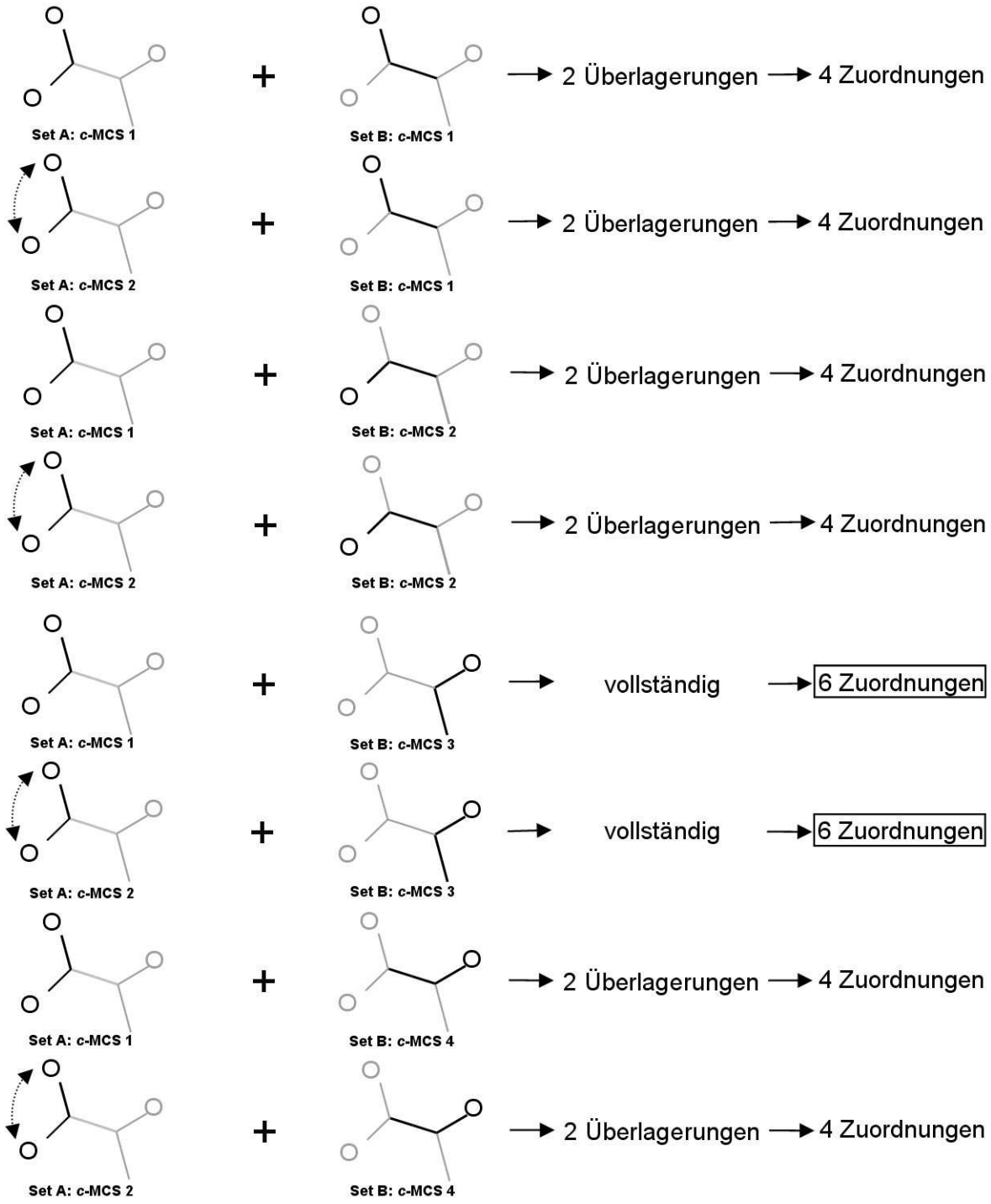


Abbildung 22: *c*-MCS-Kombinationen aus der Reaktion der Pyruvatdecarboxylase werden untersucht, um Kombinationen ohne Atomüberlagerungen zu ermitteln.

2.2.2 Rankingsystem zur Rekonstruktion komplexer Reaktionen

Die Atomzuordnung wird deutlich schwieriger, wenn mehrere Edukt- und Produktmoleküle an der Reaktion beteiligt sind. Häufig kommt es zu Überlagerungen innerhalb der *c*-MCS-Kombination, wobei zwei oder mehr *c*-MCS die gleichen Atome beanspruchen können. Um eine eindeutige Abbildung der Atome zu generieren, muss eine Entscheidung getroffen werden, welcher *c*-MCS schließlich die Atome zugeordnet werden. Bei einer manuellen Zuordnung würden größeren *c*-MCS eine höhere Priorität eingeräumt werden als Kleineren. Aus diesem Grund wurde ein Rankingsystem erarbeitet, das die *c*-MCS innerhalb einer Kombination nach ihrer Größe bewertet. Größere *c*-MCS erhalten Vorrang gegenüber den kleineren *c*-MCS. Gibt es Atomüberlagerungen zwischen größeren und kleineren *c*-MCS, so werden die Atome der größeren *c*-MCS berücksichtigt, während die Atome aus den kleineren *c*-MCS entfernt werden.

Ein Beispiel für die Funktion des Rankingsystems gibt die Reaktion der L-Arginin:Glycin-Amidinotransferase (Abbildung 23). L-Arginin und Glycin reagieren zu L-Ornithin und Guanidinacetat. Damit besitzt die Reaktion die Form A + B $\leftrightarrow$ C + D. Im ersten Schritt erfolgt ein *c*-MCS-Vergleich von jedem Eduktmolekül mit jedem Produktmolekül. Hierdurch entstehen 4 *c*-MCS Sets. Bei dieser Reaktion führt jeder Vergleich nur zu einer *c*-MCS. Hierdurch ergibt sich zunächst nur eine Möglichkeit, die *c*-MCS miteinander zu kombinieren. Die *c*-MCS dieser Kombination werden entsprechend des Rankingsystems nach ihrer Größe sortiert. Hierbei zeigt sich ein häufiger Problemfall. Die beiden *c*-MCS aus den Vergleichen von Molekül B und C, sowie Molekül B und D sind gleich groß. Beide *c*-MCS haben eine Größe von 5 Atomen und besitzen gleiche Priorität. Das automatisierte Verfahren kann zunächst nicht unterscheiden, welcher der *c*-MCS Priorität eingeräumt werden muss. Daher werden 2 Kombinationen erzeugt, wobei zunächst der einen, dann der anderen *c*-MCS Vorrang gegeben wird. Generell werden alle möglichen Permutationen erzeugt und ausgetestet, wenn eine Kombination mehrere *c*-MCS gleicher Größe enthält.

Wenn alle möglichen *c*-MCS-Kombinationen erzeugt und nach ihrer Größe sortiert wurden, werden die Atomzuordnungen schrittweise generiert (Abbildung 24). Zunächst werden die Atome der größten *c*-MCS in die Lösungsmenge aufgenommen und überlagernde Atome aus den kleineren *c*-MCS entfernt. Hierdurch können sich die Größenverhältnisse unter den verbliebenen *c*-MCS verändern. Daher erfolgt ein erneutes Umsortieren der verbliebenen *c*-MCS, bevor die Atome der zweitgrößten *c*-MCS zur Lösungsmenge hinzugefügt werden. Durch dieses iterative Verfahren werden alle *c*-MCS der Kombination betrachtet und die Atomzuordnung schrittweise vervollständigt.

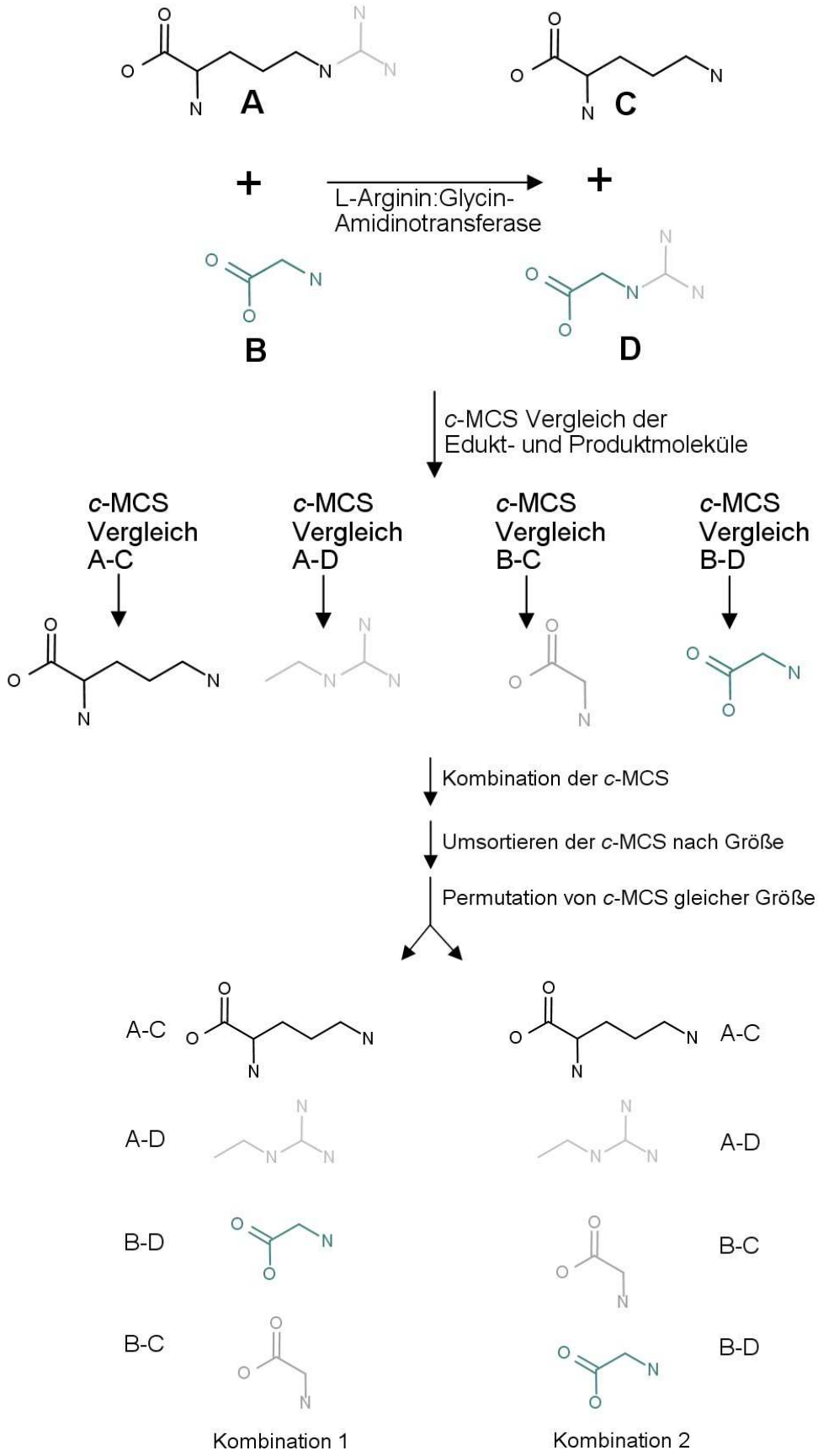


Abbildung 23: Funktion des Rankingsystems anhand des Reaktionsbeispiels der L-Arginin:Glycin-Amidinotransferase (EC-2.1.4.1).

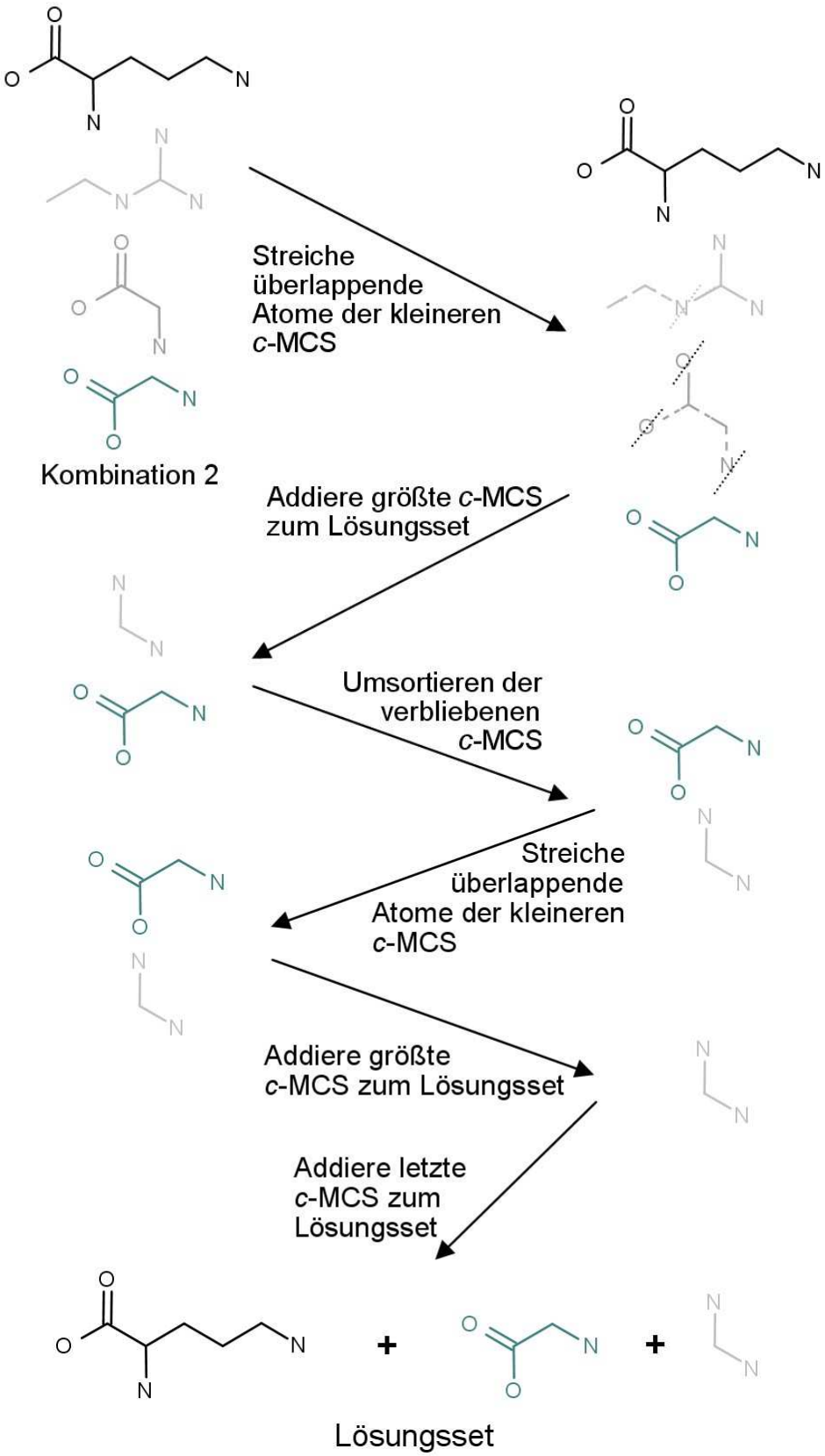


Abbildung 24: Das Schema verdeutlicht, wie die *c*-MCS der zweiten Kombination aus Abbildung 23 nacheinander in Abhängigkeit von der Größe betrachtet und zur Lösungsmenge hinzugefügt werden.

In anderen Fällen gibt es mehrere Möglichkeiten, die *c*-MCS aus den verschiedenen Sets miteinander zu kombinieren. Es werden dann alle Kombinationen betrachtet und die Lösungen abgespeichert, die zu einer maximalen Edukt/Produkt - Atomzuordnung führen. In den meisten Fällen wird durch dieses Verfahren eine vollständige Atomzuordnung generiert. Allerdings gibt es Reaktionen, die sich ohne eine Erweiterung des Verfahrens nicht nachvollziehen lassen. Weiterhin können sich die Kombinationsmöglichkeiten explosionsartig vervielfachen, wenn viele Moleküle an der Reaktion beteiligt sind. Um diese Laufzeitprobleme zu umgehen, wurden heuristische Aspekte in das Verfahren aufgenommen.

2.2.3 Erkennung lokaler Symmetrien

Der paarweise Vergleich der Moleküle führt zu einer unterschiedlichen Anzahl von *c*-MCS. Viele *c*-MCS beruhen dabei auf lokalen Symmetrien innerhalb der Moleküle. Beispielsweise besitzen in Abbildung 25 die Moleküle A und B je zwei Methylgruppen am C1-Atom. Aus graphentheoretischer Sicht ergeben sich 2 Möglichkeiten, die Methylgruppen der Moleküle einander zuzuordnen. Für die Generierung einer Atomzuordnung ist aber eine dieser äquivalenten Möglichkeiten ausreichend. Es gibt viele ähnliche Gruppen, die bei vielen Molekülen zu finden sind und die *c*-MCS-Anzahl wesentlich erhöhen können. Hierzu gehören Kohlenstoffatome, die mit Aminogruppen oder Hydroxylgruppen in Verbindung stehen, sowie Stickstoff-, Schwefel- oder Phosphatverbindungen. Es wurden daher Suchfunktionen integriert, die solche Gruppen in Molekülen auffinden und sich auf eine Atomzuordnung festlegen. Auf diese Weise werden redundante *c*-MCS aus den Sets entfernt.

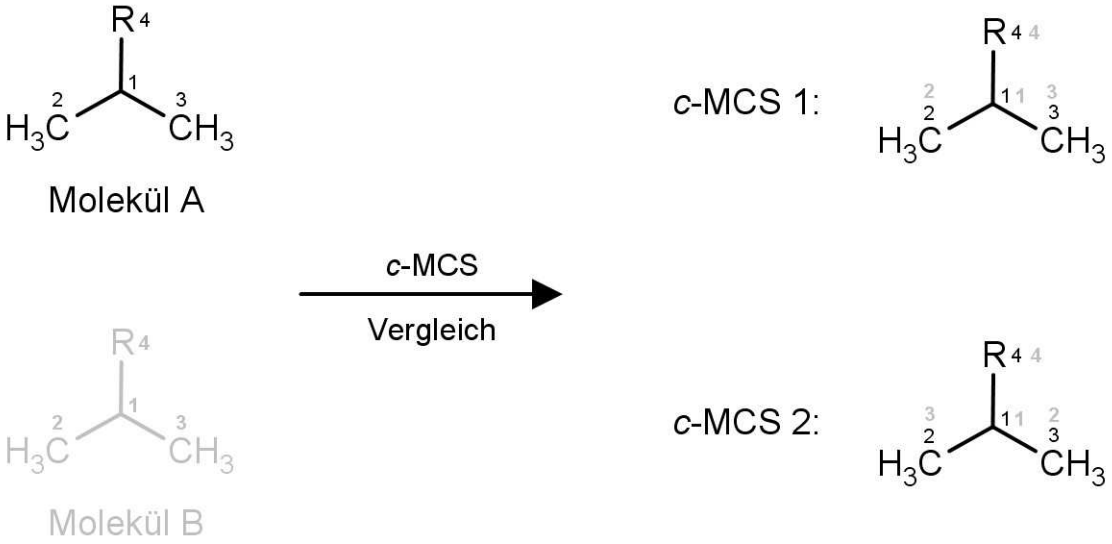


Abbildung 25: Eine lokale Symmetrie am C1-Atom führt zu äquivalenten Atomzuordnungen.

2.2.4 Atomzuordnung von Kofaktoren

An vielen enzymatischen Reaktionen sind Kofaktoren beteiligt. Hierzu gehören Energieträger wie Adenosintriphosphat (ATP) oder Oxidations- bzw. Reduktionsmittel wie beispielsweise Nicotinamidadenindinukleotid (NAD+/NADH). Einige Reaktionen benötigen diese Energie- und Reduktionsäquivalente in hoher Anzahl.

Diese Reaktionen können daher für das Atomzuordnungsprogramm ein Problem darstellen. Es muss entsprechend viele *c*-MCS-Vergleiche zwischen den Edukt- und Produktmolekülen vornehmen und die Anzahl möglicher *c*-MCS-Kombinationen kann in einem Maße zunehmen, dass die Atomzuordnung nicht mehr in akzeptabler Zeit erfolgen kann.

Andererseits sind die Veränderungen innerhalb der Kofaktoren bekannt. Sie bilden korrespondierende Molekülpaare, die sich auf der Edukt- und Produktseite auffinden lassen. Diese Eigenschaft lässt sich als heuristische Information zur Verringerung der Laufzeit verwenden. Tritt beispielsweise NAD+ als Edukt bei einer Reaktion in Erscheinung und findet sich auf der Produktseite das korrespondierende NADH Molekül, so lassen sich die Moleküle einander zuordnen und von den übrigen Molekülen der Reaktion trennen.

Korrespondierende Kofaktoren:	
NAD+	NADH
NADP+	NADPH
FAD	FADH2
FMN	FMNH2
oxidiertes Flavodoxin	reduziertes Flavodoxin
oxidiertes Ferredoxin	reduziertes Ferredoxin
Ferricytochrom	Ferrocytochrom
Ferricytochrom c	Ferrocytochrom c
Ferricytochrom c-553	Ferrocytochrom c-553
Ferricytochrom b1	Ferrocytochrom b1
Ferricytochrom b5	Ferrocytochrom b5
Ferricytochrom b-561	Ferrocytochrom b-561
Fe(III)	Fe(II)
ATP + H_2O	ADP + Phosphat
ATP + H_2O	AMP + Diphosphat

Tabelle 3: Auflistung der berücksichtigten Kofaktor Moleküle

Tabelle 3 enthält eine Auflistung der korrespondierenden Kofaktoren. Werden entsprechende Molekülpaare entdeckt, werden die Moleküle durch *c*-MCS-Suche verglichen und die zugehörigen Edukt- und Produktatome zugeordnet. Anschließend werden die Moleküle maskiert und dem weiteren Molekülvergleich entzogen.

2.2.5 Bron-Kerbosch-Algorithmus

Der *c*-MCS-Algorithmus und das Rankingsystem erlauben in den meisten Fällen eine vollständige Atomzuordnung aller Edukt- und Produktatome. Bei einigen Reaktionen reicht dieses Verfahren allerdings nicht aus, um die Atomzuordnung abzuschließen. Problematisch sind Fälle, wo eine besonders kleine Struktur von einem Molekül auf ein anderes übertragen wird, denn die übertragene Struktur muss nicht Bestandteil der *c*-MCS beider Moleküle sein. Ein Beispiel für diese Situation gibt die Reaktion der Amino-acid N-Acetyltransferase (Abbildung 26). Dieses Enzym katalysiert die Übertragung einer Acetylgruppe von Acetyl-CoA auf L-Glutamat. Der *c*-MCS-Vergleich von L-Glutamat mit N-Acetyl-L-Glutamat ergibt eine *c*-MCS, die dem L-Glutamat Molekül entspricht, da die Wasserstoffatome auf dieser Ebene noch nicht berücksichtigt werden. Ebenso liefert der Vergleich von Actyl-CoA mit CoA eine *c*-MCS, die mit dem CoA Molekülteil übereinstimmt.

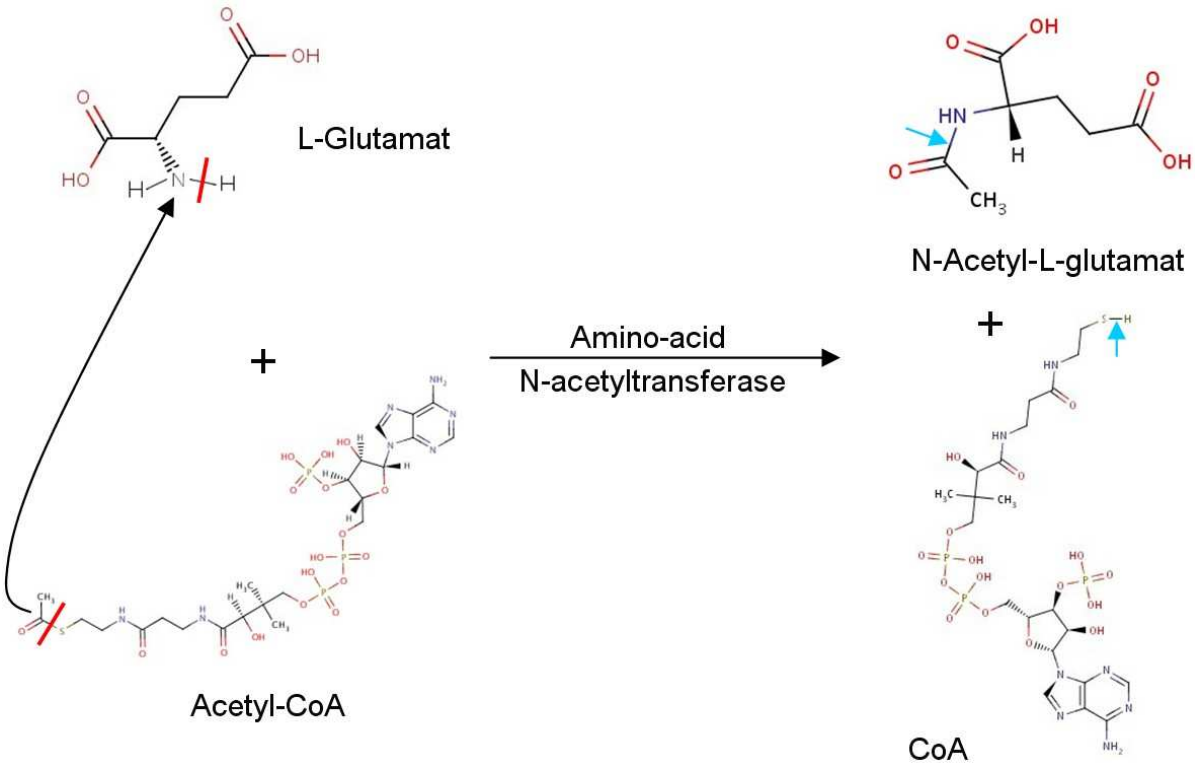


Abbildung 26: Die Amino-acid N-Acetyltransferase (EC-2.3.1.1) ist ein Beispiel für einen Reaktionstyp, wo das Ranking-System nicht zu einer vollständigen Atomzuordnung führt.

Diese beiden Strukturen werden durch das Rankingsystem zugeordnet. Nicht zugeordnet wird hingegen die Acetylgruppe, die auf das L-Glutamat-Molekül übertragen wird. Ursache hierfür ist, dass der *c*-MCS-Vergleich von Acetyl-CoA und N-Acetyl-L-glutamat zu einer *c*-MCS führt, die nicht die Acetylgruppe des Acetyl-CoA beinhaltet. Vielmehr weist ein Bereich innerhalb des CoA-Molekülteils größere Ähnlichkeit mit N-Acetyl-L-glutamat auf.
Als Lösung für diesen Problemtyp wurde eine Aufteilung des Molekülvergleichs in mehrere Ebenen gewählt. Führt der *c*-MCS-Vergleich und das Rankingsystem zu unvollständigen Atomzuordnung, werden die verbliebenen Strukturen mit Hilfe einer weiteren Bron-Kerbosch-Variante verglichen. Diese Variante entspricht weitgehend dem ursprünglichen Bron-Kerbosch-Algorithmus (Abschnitt 2.1.3).

2.2.6 Vergleich von kleinen Molekülen durch den Bron-Kerbosch-Algorithmus

Anders als der schnellere *c*-MCS-Algorithmus findet der ursprüngliche Bron-Kerbosch-Algorithmus auch nicht-zusammenhängende MCS. Dies ist besonders hilfreich, wenn mehrere kleine Fragmente oder einzelne Atome auf der Edukt- und Produktseite zurückbleiben.
Bei besonders kleinen Molekülen bietet es an, auf eine *c*-MCS-Suche auf der ersten Molekülvergleichsebene zu verzichten und diese erst mit Hilfe des Bron-Kerbosch-Algorithmus nachträglich zuzuordnen. Denn der Vergleich von sehr kleinen mit größeren Molekülen kann zu einer hohen Anzahl von *c*-MCS führen. Beispielsweise entstehen 40 *c*-MCS, wenn Methan mit einem größeren Molekül mit 40 Kohlenstoffatomen verglichen wird. Hierdurch kann sich die Anzahl der Kombinationen erheblich vervielfachen, wenn die *c*-MCS der verschiedenen Vergleichsets kombiniert werden. Ein nachträgliches Zuordnen mit Hilfe des Bron-Kerbosch-Algorithmus führt hier zu einer wesentlichen Verbesserung der Laufzeit.

2.2.7 Erkennung von aromatischen Ringen

Aromaten sind eine Molekülklasse, die planare und zyklische Ringsysteme mit konjugierten Doppelbindungen enthalten (Hart, 2002). Alle Atome des Ringsystems sind sp2-hybridisiert. Dies bedeutet, dass sich 2 Elektronen der Doppelbindungen auf besonders günstigen Energieniveaus befinden. Sie werden als π-Elektronen bezeichnet und sind nicht in den π-Orbitalen zwischen den beiden C-Atomen lokalisiert, sondern verteilen sich in einer Elektronenwolke über die Atome des gesamten aromatischen Systems. Die Wolke bildet ein so genanntes π-System aus, das sich oberhalb und unterhalb des Ringsystems befindet. Diese Überlappung der π-Orbitale wird auch als

Konjugation bezeichnet und verleiht Aromaten spezifische chemische und physikalische Eigenschaften. Hierdurch unterscheiden sich Aromaten von den übrigen organischen Verbindungen, den Aliphaten.

Aromatische Ringe werden häufig nicht gekennzeichnet. Wenn gleich die π-Elektronen delokalisiert vorliegen und sich nicht fest einer Bindung zuordnen lassen, werden aromatische Ringe zumeist als Ringsysteme mit alternierenden Doppelbindungen angegeben. Die Zuordnung der Einfach- und Doppelbindungen erfolgt willkürlich und bleibt dem Zeichner des Moleküls vorbehalten. Die in den „Molfiles“ kodierten Moleküle werden zunächst auch gezeichnet und durch Programme, wie das Programm MDL ISIS draw, in das Molfile-Format übersetzt. Somit können Doppelbindungen von aromatischen Ringen in korrespondierenden Edukt- und Produktmolekülen unterschiedlich eingezeichnet sein. Die Folge ist in diesem Fall, dass später für den Ring ein Elektronentransfermuster errechnet wird, obwohl keine Änderungen in der Elektronenverteilung stattfinden.

In manchen Fällen bilden aromatische Ringe symmetrische Strukturen. Hierdurch können Atomzuordnungen entstehen, die zu falschen Elektronentransfermustern führen. Ein Beispiel hierfür ist die Reaktion der Methylenetetrahydrofolat-Reduktase (Abbildung 27), in der 5-Methyltetrahydrofolat zu 5,10-Methylenetetrahydrofolat umgewandelt wird. In der Molekülmitte befindet sich ein aromatischer Ring, der aus der Reaktion unverändert hervorgeht. Der *c*-MCS-Algorithmus berücksichtigt keine Doppelbindungen. Daher stellt der aromatische Ring in dem Molekül eine lokale Symmetrie dar. Wie in Abbildung 28 dargestellt, ergeben sich 2 Möglichkeiten die Ringatome der beiden Moleküle einander zuzuordnen. Bei Möglichkeit A korrespondieren die Atomzuordnungen mit den Doppel- und Einfachbindungen. Die Doppel- und Einfachbindungen sind einander zugeordnet und der Ring bleibt unverändert. Anders bei der zweiten Möglichkeit (B): Hier werden die Atome so zugeordnet, dass die Doppelbindungen an die Stelle der Einzelbindungen treten und umgekehrt. Bei dieser Atomzuordnung wird später ein Elektronenaustausch errechnet, wobei 3 zusätzliche Bindungen entstehen und 3 Bindungen gespalten werden.

Um derartige falsche Zuordnungen auszuschließen, wurde ein Algorithmus zur Erkennung von aromatischen Ringen entwickelt. Dieser Algorithmus wurde in C++ implementiert und als Klasse „Aromatic_Rings“ in das Programm integriert. Der Algorithmus beruht auf einer Backtrackingsuche. Ausgehend von jeder Doppelbindung eines Moleküls wird zunächst nach einer alternierenden Folge von Einfach- und Doppelbindungen gesucht. Diese Folge von Bindungen kann dann zum Ring geschlossen werden, wenn der Algorithmus nach einigen Schritten wieder zu der Ausgangsbindung gelangt. Die gefundene Ringstruktur wird dann auf die Aromatizitätskriterien untersucht.

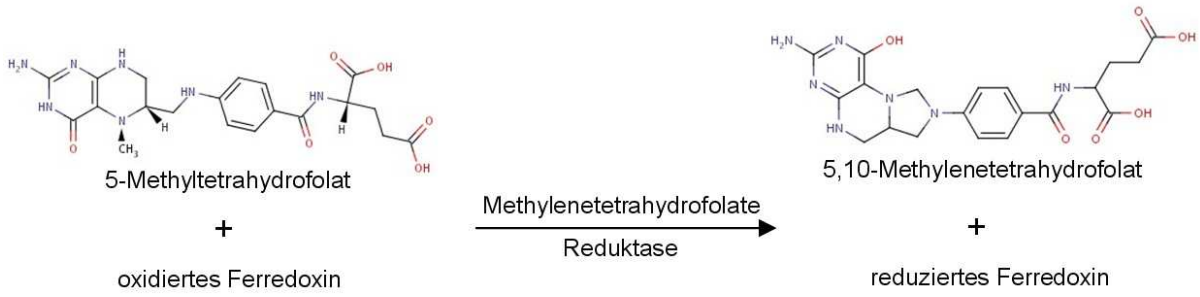


Abbildung 27: Reaktion der Methylenetetrahydrofolat-Reduktase (EC-Nummer 1.5.7.1).

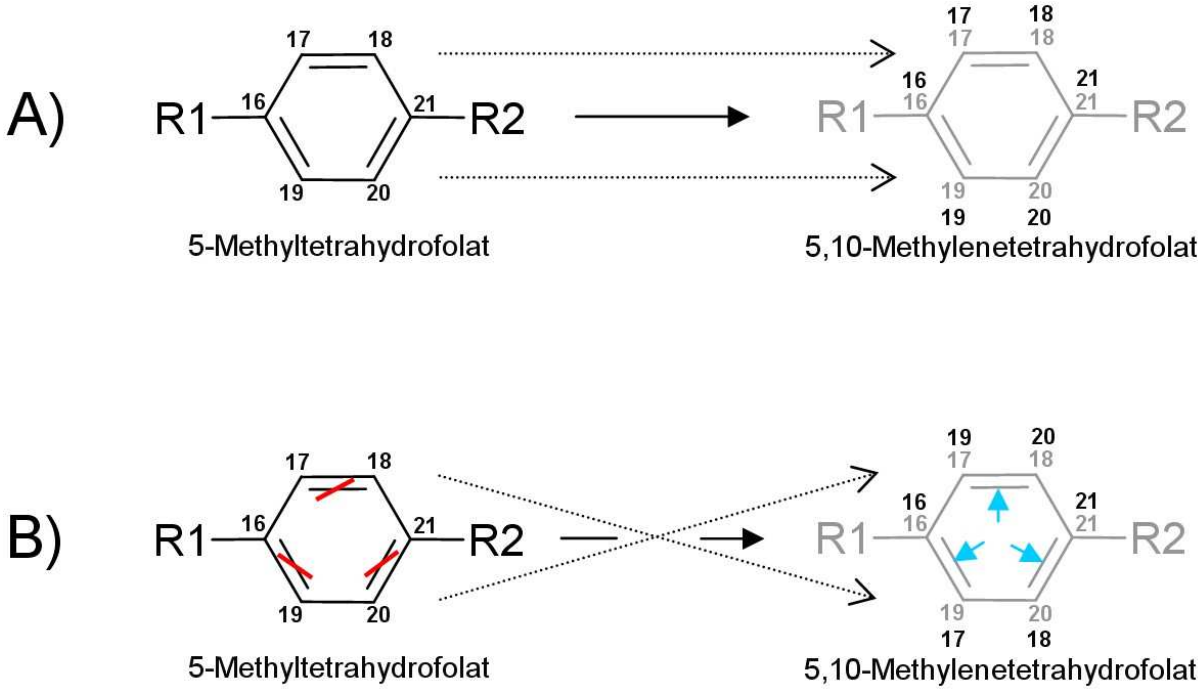


Abbildung 28: Der c-MCS-Algorithmus unterscheidet nicht zwischen Doppel- und Einfachbindungen. Es gibt daher 2 Möglichkeiten, den Benzolring der Moleküle 5-Methyltetrahydrofolat und 5,10-Methylenetetrahydrofolat aufeinander abzubilden. Bei der Atomzuordnung der zweiten Möglichkeit (B) entsteht vermeintlich ein Elektronenaustausch.

Nicht alle aromatischen Ringe beinhalten ausschließlich ein System von alternierenden Doppel- und Einfachbindungen. Es gibt auch aromatische Ionen, wo auf eine Einfachbindung keine Doppelbindung folgt, sondern ein positiv oder negativ geladenes Kohlenstoffatom oder Heteroatom. Entsprechende Heteroatome können Stickstoff-, Sauerstoff- oder Schwefelatome sein. Hierzu gehören beispielsweise die kleineren Cyclopropenium-Salze, die in ihrem Ring nur 2 π-Elektronen

und ein positiv geladenes Kohlenstoffatom enthalten. Positiv geladene Ringe findet man aber auch bei Derivaten des Cycloheptatrien oder man muss mit einer negativen Ladung bei Cyclopentadien-Derivaten rechnen. Der Algorithmus prüft auch diese Möglichkeiten. Wenn auf eine Doppelbindung eine Einfachbindung mit einem geladenen Atom folgt, so wird im nächsten Schritt wieder nach einer Einfachbindung mit einer darauf folgenden Doppelbindung gesucht.

Neben den aromatischen Ionen, gibt es zahlreiche weitere Ausnahmen, wo freie Elektronenpaare von Heteroatomen am Aufbau des π-Systems beteiligt sind. Aus diesem Grund muss die Anzahl der Bindungen, die von den Heteroatomen ausgehen, ermittelt werden. Bindungsanzahl und Ladung eines Heteroatoms gibt Aufschluss darüber, ob das Heteroatom freie Elektronenpaare besitzt, die es dem π-System zur Verfügung stellen kann. Aromaten dieses Typs sind beispielsweise Furan, Pyrrol oder Thiophen, um nur einige zu nennen.

Gefundene Ringsysteme mit alternierenden Doppel- und Einfachbindungen, mit Ladungen, oder Heteroatomen mit freien Elektronenpaaren werden auf die Hückel-Regel geprüft. Danach sollen die delokalisierten Elektronen des konjugierten Elektronensystems einer bestimmten Anzahl entsprechen, die sich durch Formel (4n + 2) errechnen lässt. Dabei gibt „n“ die Anzahl der Ringsysteme an. Entsprechend müssen zunächst 2 (Cyclopropen-Derivate) oder 6 delokalisierte Elektronen in einem Ring vorhanden sein.

Das Auffinden von größeren konjugierten Elektronensystemen mit mehreren verbundenen Ringen erfordert ein komplexeres Suchverfahren. Ein Beispiel zeigt Abbildung 29, wo 2 aromatische Ringe miteinander verbunden sind. Das konjugierte Elektronensystem enthält 10 delokalisierte Elektronen und entspricht somit der Hückel-Regel. Ring A lässt sich mit Hilfe des oben beschriebenen Algorithmus auffinden. Der zweite Ring B hingegen nicht, denn die beiden Doppelbindungen des Rings werden durch 3 Einfachbindungen voneinander getrennt. Ring B lässt sich daher nur auffinden, wenn Ring A bereits als aromatischer Ring identifiziert wurde. Denn in diesem Fall kann die gemeinsame Bindung der beiden Ringe als Doppelbindung gewertet werden.

Die Suche nach verbundenen aromatischen Ringen erfolgt daher in Form eines iterativen Verfahrens. Zunächst werden mit Hilfe des beschriebenen Backtracking-Algorithmus Ringe ausfindig gemacht, die alleine den Aromatizitätskriterien entsprechen. Die Bindungen dieser Ringe werden in der Folge als Doppelbindungen gewertet. Anschließend erfolgt eine neue Suche ausgehend von den Doppelbindungen, die bislang noch keinem aromatischen Ring zugeordnet werden konnten. Auf diese Weise wird nun auch Ring B als aromatisch identifiziert. Neben Ring A und B wären weitere aromatische Ringe denkbar, die mit dem bisherigen Ringsystem in Verbindung

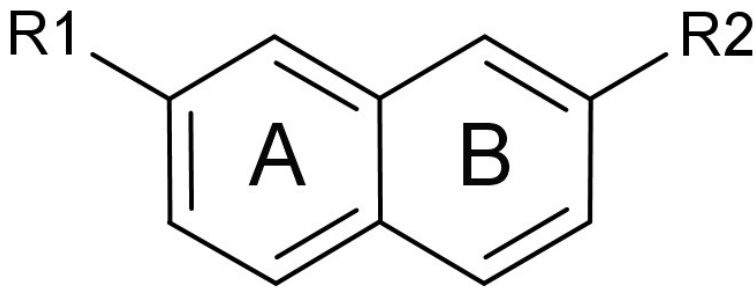


Abbildung 29: Ring B lässt sich nur in Verbindung mit Ring A als aromatisch identifizieren.

stehen könnten. Deshalb fährt das iterative Verfahren solange fort, bis für alle verbliebenen Doppelbindungen definitiv ausgeschlossen werden kann, dass sie einem konjugierten System angehören.
Neben den hier beschriebenen Aspekten für die Suche nach Aromaten, lassen sich zahlreiche komplexere Fälle konstruieren, die man in die Suche einbeziehen müsste. Da diese Fälle in der Praxis bei biochemischen Molekülen aber nicht in Erscheinung treten, werden sie durch den Suchalgorithmus nicht berücksichtigt (z.B. Pyrolringe in Cytochromen).
Obwohl sich der Algorithmus eines Backtrackingverfahrens bedient, werden die aromatischen Ringsysteme aller Moleküle ohne Laufzeitprobleme ermittelt. Damit die Suche nicht mehrere Ringe durchläuft, wird die Tiefensuche nach wenigen Schritten abgebrochen. Hiermit ermöglicht es der Algorithmus, korrespondierende aromatische Ringe der Edukt- und Produktseite einander zuzuordnen. Wenn ein Ring vor als auch nach der Reaktion die Kriterien für einen aromatischen Ring erfüllt, können nicht-korrekte Bindungsaustausche erkannt und ausgeschlossen werden.

2.2.8 Addition und Zuordnung der Wasserstoffatome

Als Eingabe für das Programm werden so genannte Molfiles verwendet. Diese Textfiles enthalten die Informationen über die Atome und Bindungen eines Moleküls. Dabei werden aber nur die Atome und Bindungen der schwereren Atome des Molekülrückgrates beschrieben. Die kleineren Wasserstoffatome sind hingegen nur in Ausnahmefällen in den Molfiles enthalten. Eine Ausnahme ist beispielsweise molekularer Wasserstoff (H_2). Wasserstoffatome können nur eine Bindung eingehen und spielen beim Aufbau einer verzweigten Molekülstruktur eine untergeordnete Rolle. Den Kohlenstoffatomen und Heteroatomen kommt hingegen eine größere Bedeutung zu. Diese Atome bilden das Molekülrückgrat. Die Wasserstoffatome verteilen sich entlang der Rückgratatome. Daher lassen sich die Wasserstoffatome mit Hilfe der Rückgratatome herleiten. Die Anzahl der Wasserstoffatome eines Kohlenstoff- oder Heteroatoms ist abhängig von seinem

Elementtyp, seiner Ladung und der Anzahl seiner Bindungen zu den anderen Atomen des Molekülrückgrats.

Das im Rahmen dieser Arbeit entwickelte Programm berücksichtigt zunächst nur die größeren Atome aus den Molfiles für die Generierung einer Atomzuordnung der Eduktatome auf die Produktatome. Erst wenn dieser Vorgang abgeschlossen ist, werden die Wasserstoffatome zu den Molekülen addiert, um die Atomzuordnung zu vervollständigen.

Abbildung 30 zeigt, welche Wasserstoffatome für das Molekül 4-Aminobutanal bestimmt werden. Das Molfile enthält nur die 6 Atome des Molekülrückgrats. Dieses besteht aus 4 Kohlenstoff-, einem Sauerstoff- und einem Stickstoffatom. Kohlenstoff gehört der vierten Hauptklasse an und besitzt 6 Elektronen. Die vier äußeren Elektronen der zweiten Schale können je eine Bindung eingehen, so dass in organischen Molekülen von Kohlenstoffatomen in der Regel vier Bindungen ausgehen. Die Kohlenstoffatome C2, C3 und C4 besitzen bereits zwei Bindungen. Da diese Kohlenstoffatome keine Ladung besitzen, kann davon ausgegangen werden, dass noch zwei Bindungen zu Wasserstoffatomen fehlen. Diese Bindungen werden zusammen mit den Wasserstoffatomen ergänzt. Kohlenstoffatom C5 verfügt bereits über drei Bindungen. Daher wird eine Bindung mit einem Wasserstoffatom hinzugefügt.

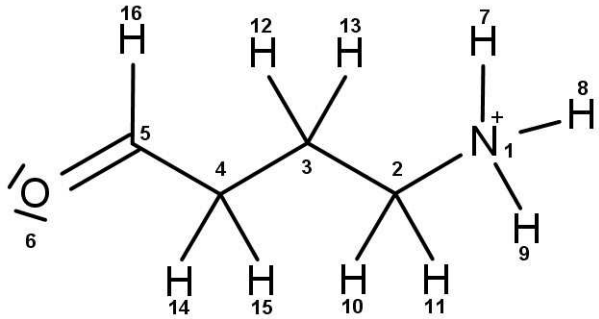


Abbildung 30: Molekül 4-Aminobutanal. Die Wasserstoffatome sind nicht in den „Molfiles“ kodiert. Sie werden mit Hilfe der Kohlenstoff- und Heteroatome hinzugefügt.

Stickstoff gehört der fünften Hauptklasse an und besitzt 5 Elektronen auf der äußersten Schale. Hierdurch entsteht eine $2s^2p^3$ Elektronenkonfiguration. Zur Oktett-Komplettierung kann Stickstoff demnach drei kovalente Bindungen eingehen. Die restlichen beiden Elektronen liegen als freies Elektronenpaar vor und können als Base agieren. Wenn das freie Elektronenpaar ein H^+ bindet, können von dem Stickstoffatom auch vier Bindungen ausgehen. Das Stickstoffatom wird hierdurch positiv geladen. In dieser Form liegt auch das Stickstoffatom von 4-Aminobutanal vor.

Sauerstoff besitzt 8 Elektronen, von denen sich 6 auf die äußere Schale verteilen. Das 2s-Orbital und ein 2p-Orbital werden vollständig mit Elektronen gefüllt. Diese Elektronen treten als zwei freie

Elektronenpaare in Erscheinung und tragen nichts zur Bindung bei. Die Bindungseigenschaften des Sauerstoffs werden von 2 Valenzelektronen aus den 2p-Orbitalen bestimmt. Hierdurch ist Sauerstoff in der Lage, 2 Bindungen einzugehen. Das Sauerstoffatom 6 von 4-Aminobutanal besitzt keine Ladung und ist mit dem Kohlenstoffatom 5 über zwei Bindungen verbunden. Hier können keine Wasserstoffatome hinzugefügt werden. Anders ist dies beispielsweise bei Wassermolekülen. Das Molfile von Wasser enthält nur ein ungeladenes Sauerstoffatom. Hier werden zwei Wasserstoffatome mit entsprechender Bindung hinzugefügt. In dieser Weise erfolgt auch das Hinzufügen der Wasserstoffatome bei den anderen Heteroatomen organischer Verbindungen.
Wenn bei allen Molekülen der Reaktion die Wasserstoffatome hinzugefügt wurden, erfolgt die Zuordnung der Eduktwasserstoffatome auf die Produktwasserstoffatome. Abbildung 31

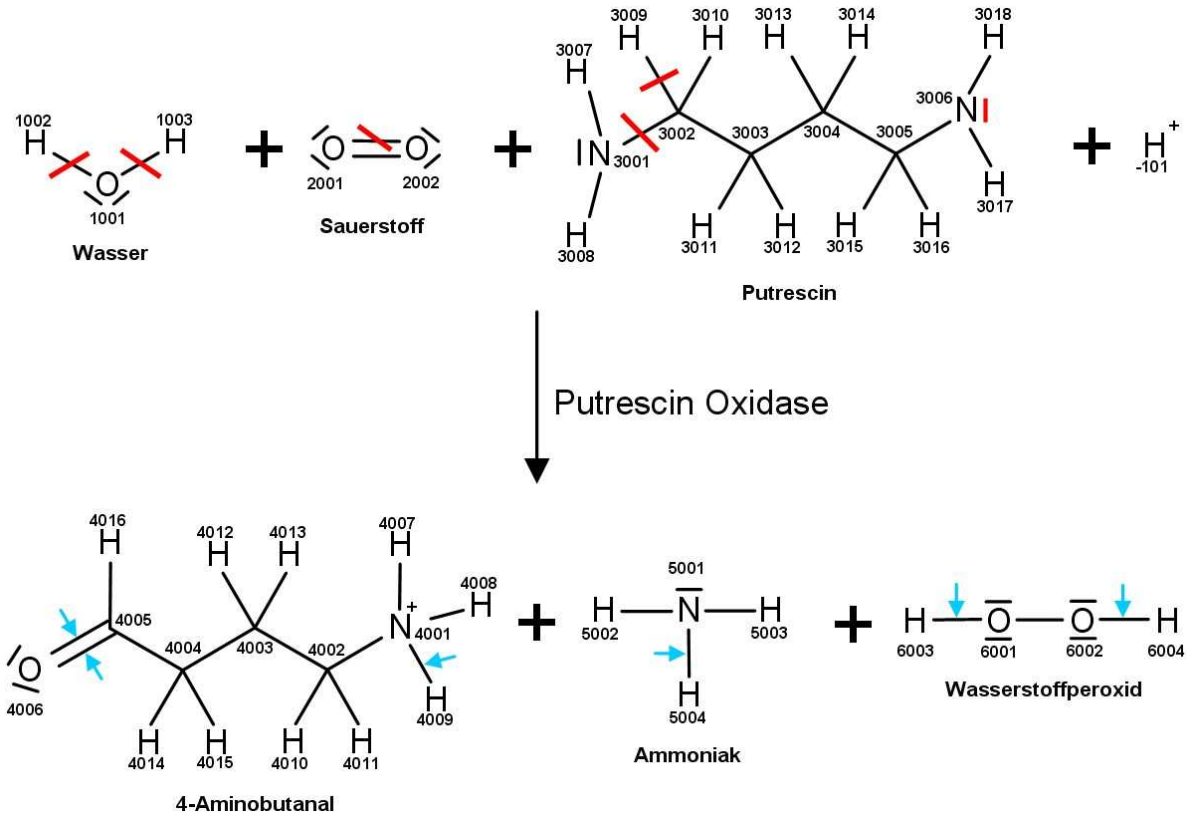


Abbildung 31: Reaktion der Putrescin-Oxidase (EC-Nummer 1.4.3.10). Alle Atome besitzen eine eindeutige Nummerierung, um Verwechslungen bei der Atomzuordnung zu vermeiden.

zeigt die Reaktion der Putrescin-Oxidase. Die Kohlenstoff- und Heteroatome des Molekülrückgrats wurden bereits durch die *c*-MCS-Kombination zugeordnet. Diese Atomzuordnung wird durch das Programm schrittweise durchlaufen und dabei die Wasserstoffatome zugeordnet. Ist beispielsweise ein Kohlenstoffatom sowohl vor als auch nach der Reaktion mit zwei Wasserstoffatomen verbunden, so können diese einander zugeordnet werden. Die Zuordnung dieser zwei Wasserstoffatome erfolgt willkürlich. Sie werden auf zweidimensionaler Ebene betrachtet und unterscheiden sich nicht in den Eigenschaften.

In einigen Fällen stimmt die Anzahl der Wasserstoffatome eines Rückgratatoms auf der Edukt- und Produktseite nicht überein. Beispielsweise besitzt das Kohlenstoffatom 3002 auf der Eduktseite zwei Wasserstoffatome. Auf der Produktseite besitzt es als Atom 4005 dagegen nur ein Wasserstoffatom. Offensichtlich wurde das Wasserstoffatom im Laufe der Reaktion abgespalten. In anderen Fällen nimmt die Anzahl der Wasserstoffatome zu. In dieser Reaktion besitzen die Atome der Zuordnungen 2001:6001, 2002:6002, 3001:5001, 3006:4001 auf der Produktseite mehr Wasserstoffatome als auf der Eduktseite. Diese Atome gehen neue Bindungen mit Wasserstoffatomen ein.

Die Wasserstoffatome, die auf der Eduktseite abgespalten werden oder auf der Produktseite neue Bindungen eingehen, werden auf diese Weise identifiziert und gespeichert. Wenn der Vergleich der Wasserstoffatome abgeschlossen ist, werden die abgespaltenen Edukt-Wasserstoffatome den Wasserstoffatomen auf der Produktseite zugeordnet, die dort eine neue Bindung eingehen. In vielen Fällen kann hier eine 1:1 Zuordnung erfolgen ohne überzählige Wasserstoffatome.

Allerdings wurde mit der Putrescin-Oxidase eine Reaktion gewählt, bei der die Anzahl der Wasserstoffatome nicht übereinstimmt. Auf der Produktseite werden 4 Wasserstoffatome neu gebunden, während nur 3 Wasserstoffatome auf der Eduktseite abgespalten werden. Auch insgesamt stehen 14 Wasserstoffatomen auf der Eduktseite 15 Wasserstoffatome auf der Produktseite gegenüber. Der Reaktion liegt kein Fehler zugrunde. Vielmehr benötigt die Reaktion ein H^+-Ion auf der Eduktseite. Bei vielen Reaktionen treten H^+-Ionen auf der Edukt- oder Produktseite in geringer oder hoher Anzahl auf. Das Programm erkennt fehlende H^+-Ionen und fügt sie dem Reaktionssystem hinzu. Alle H^+-Ionen erhalten als Nummerierung die negative Zahl „-101“, um sie eindeutig von den übrigen Atomen des Reaktionssystems zu unterscheiden.

2.3 R-Matrix-Berechnung

Das Generieren der Atomzuordnung ist der aufwendigste Teil der automatisierten R-Matrix-Berechnung. Liegt eine 1:1 Zuordnung der Eduktatome auf die Produktatome vor, lassen sich die Eduktmatrix (B) und die Produktmatrix (E) erstellen und die Reaktionsmatrix (R) anhand von Gleichung 2 berechnen:

$$R = E - B \tag{2}$$

Bei der Reaktion der Pyruvat-Decarboxylase werden zunächst die Metabolite durch den *c*-MCS-Algorithmus verglichen und eine atomare Zuordnung generiert (Abbildung 32).

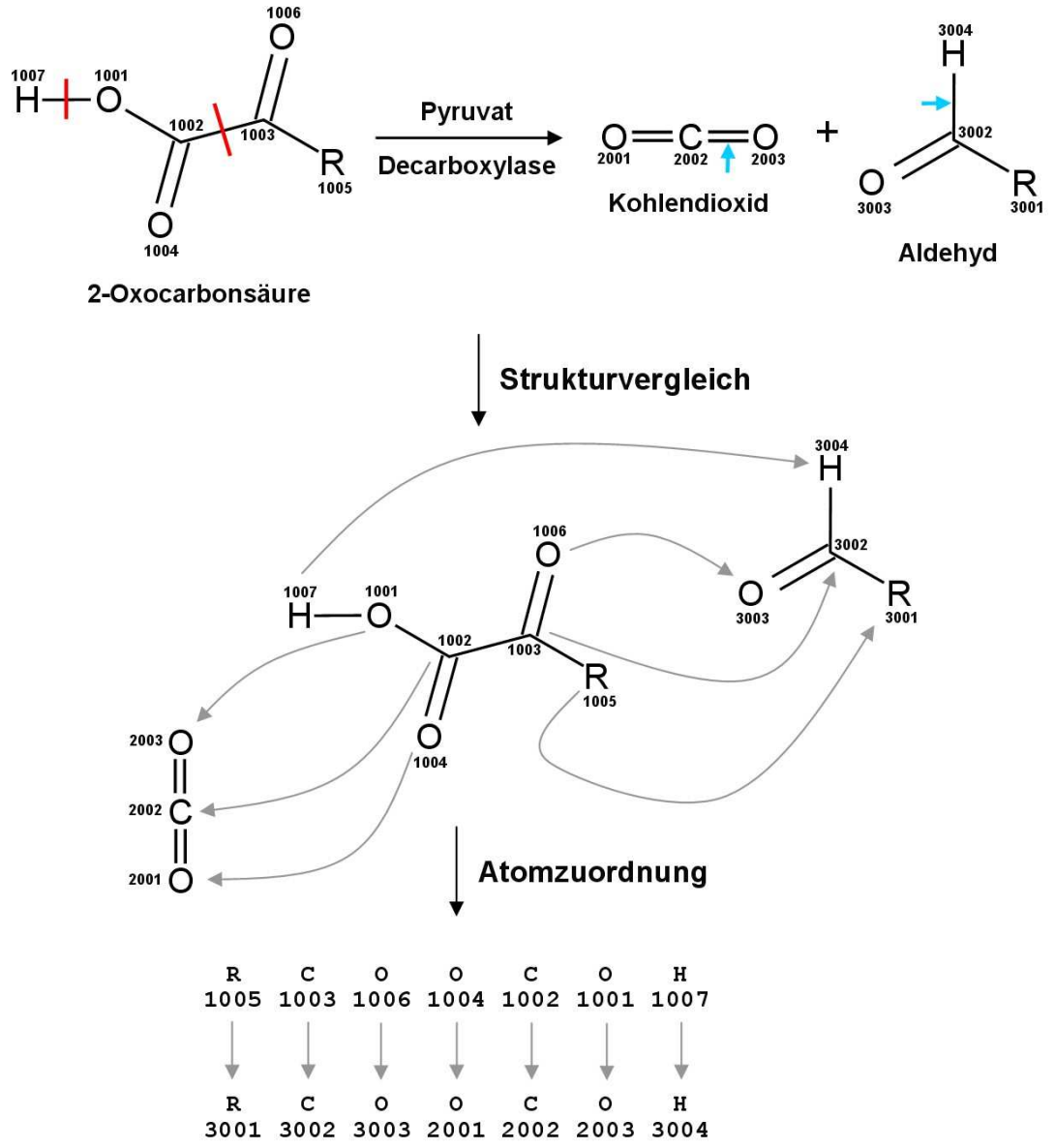


Abbildung 32: Strukturvergleich und Atomzuordnung für die Metabolite der Pyruvat-Decarboxylase (EC-Nummer 4.1.1.1).

Produkt Matrix E

-	R	C	O	O	C	O	H
	3001	3002	3003	2001	2002	2003	3004
R 3001	0	1	0	0	0	0	0
C 3002	1	0	2	0	0	0	1
O 3003	0	2	4	0	0	0	0
O 2001	0	0	0	4	2	0	0
C 2002	0	0	0	2	0	2	0
O 2003	0	0	0	0	2	4	0
H 3004	0	1	0	0	0	0	0

-

Edukt Matrix B

-	R	C	O	O	C	O	H
	1005	1003	1006	1004	1002	1001	1007
R 1005	0	1	0	0	0	0	0
C 1003	1	0	2	0	1	0	0
O 1006	0	2	4	0	0	0	0
O 1004	0	0	0	4	2	0	0
C 1002	0	1	0	2	0	1	0
O 1001	0	0	0	0	1	4	1
H 1007	0	0	0	0	0	1	0

||

Reaktionsmatrix R

0	0	0	0	0	0	0
0	0	0	0	-1	0	1
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	-1	0	0	0	1	0
0	0	0	0	1	0	-1
0	1	0	0	0	-1	0

Abbildung 33: Die R-Matrix der Pyruvat-Decarboxylase ergibt sich aus der Substraktion der Eduktmatrix von der Produktmatrix. Die Matrixwerte errechnen sich nach der Gleichung $r_{ij} = e_{ij} – b_{ij}$, wobei i und j die Indices der Matrix repräsentieren.

Anschließend werden die Edukt- und Produktmatrix generiert und die Reaktionsmatrix errechnet (Abbildung 33). Dabei nimmt jedes Atom der Reaktion innerhalb der Matrizen von B, E und R den gleichen Index ein. Um jedes Atom der verschiedenen Moleküle unterscheiden zu können, erhält

jedes Atom durch das Programm eine eindeutige Nummerierung. Durch die Molfiles wird eine Nummerierung vorgegeben, die aber immer mit der Zahl 1 beginnt und daher für ein Set von Molekülen nicht eindeutig ist. Daher wird bei jedem Molekül zu der Atomnummer des Molfiles ein Wert hinzuaddiert. Ein Wert von Tausend wird zu den Atomnummern des ersten Moleküls addiert, das dem Programm übergeben wird. Bei dem zweiten Molkül wird ein Wert von Zweitausend und bei dem Dritten ein Wert von Dreitausend zu den Atomnummern der Molefiles addiert.
Die R-Matrix gibt Aufschluss darüber, welche Bindungen entstehen oder gespalten werden. Positive Einträge sehen für Bindungen, die gebildet werden. Negative Werte repräsentieren hingegen Bindungen, die gespalten werden. Die R-Matrix der Pyruvat-Decarboxylase enthält insgesamt vier positive und vier negative Einträge. Damit werden auf der Eduktseite 2 Bindungen gespalten, während auf der Produktseite 2 Bindungen neu entstehen. Die Summe aller Matrixeinträge in R ergibt den Wert Null.
Aus Abbildung 33 wird ersichtlich, dass die Reaktionsmatrix Zeilen und Spalten enthält, die nur aus Null-Einträgen bestehen. Die korrespondierenden Atome sind nicht an der Reaktion beteiligt. Bei größeren Eingangsmolekülen sind die Matrixdimensionen groß, obwohl nicht mehr Atome an der Reaktion beteiligt sein müssen. Die redundanten Nullzeilen und Spalten werden aus den R-Matrizen entfernt. Die verkleinerte R-Matrix wird zusammen mit den Atomzuordnungen ausgegeben. Die erste Zeile und Spalte enthalten die Eduktatome. Diesen Atomen zugeordnet befinden sich darunter und daneben die korrespondierenden Produktatome (Abbildung 34).
Die Zeile und Spalte eines Atoms lassen sich innerhalb der R-Matrix beliebig verschieben. Um dennoch einen effizienten Vergleich von R-Matrizen zu ermöglichen, werden sie mit Hilfe eines Verfahrens zur Kanonisierung umgeformt. Das Kanonisierungsverfahren wird in Abschnitt 2.4 beschrieben.

```
-                    C    C    O    H
                     1003 1002 1001 1007
                     C    C    O    H
                     3002 2002 2003 3004

 C 1003  C 3002   0   -1    0    1
 C 1002  C 2002  -1    0    1    0
 O 1001  O 2003   0    1    0   -1
 H 1007  H 3004   1    0   -1    0
```

Abbildung 34: Ausgegebene R-Matrix der Pyruvat-Decarboxylase.

2.3.1 R-Matrix-Berechnung von Isomerasereaktionen

Die Enzyme der 5. Hauptklasse katalysieren geometrische oder strukturelle Änderungen eines Moleküls. Dabei entsteht eine isomere Verbindung, welche die gleiche Summenformel besitzt. Dies bedeutet, dass sich durch die Reaktion weder die Anzahl der Atome noch die Elementzusammensetzung des Ausgangsmoleküls verändert.

Einige Isomerasereaktionen lassen sich mit Hilfe des Dugundji-Ugi-Modells beschreiben. Hierzu gehören die intramolekularen Oxidoreduktasen (Subklasse EC-5.3) oder die intramolekularen Transferasen (Subklasse EC-5.4). Die Reaktionen der Racemasen lassen sich hingegen nicht mit dem Dugundji-Ugi-Modell beschreiben. Diese Enzyme katalysieren eine Veränderung in der Stereoisomerie. Die Substituenten eines chiralen Atoms werden umgelagert. Hierdurch verändert sich die optische Aktivität (Optische Isomerie) und die Wechselwirkungen zu anderen chiralen Molekülen. Die Konnektivität zwischen den Atomen innerhalb des Edukt- und Produktmoleküls ändert sich hingegen nicht. Die Alanin-Racemase (EC-Nummer 5.1.1.1) katalysiert die Umsetzung von L-Alanin zu D-Alanin.

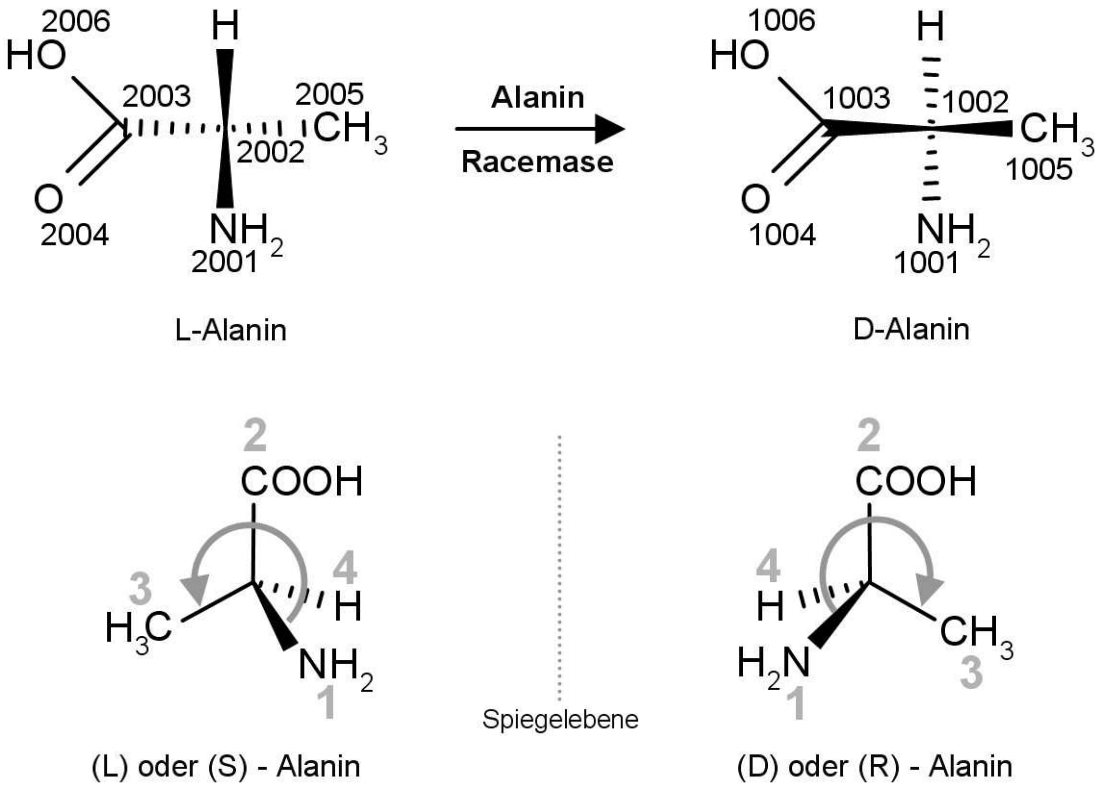


Abbildung 35: Substrat und Produkt der Alanin-Racemase verhalten sich wie Objekt und Spiegelbild. Die Konnektivität der beiden Moleküle ist identisch. Beide Formen drehen aber die Schwingungsebene von linear polarisiertem Licht im Uhrzeigersinn (D-Form) oder dagegen (L-Form).

Wie in Abbildung 35 gezeigt, verhalten sich beide Moleküle wie Bild und Spiegelbild (Spiegelbild-Isomerie) und sind durch Drehung nicht ineinander überführbar. Solche Moleküle werden als Enantiomere bezeichnet. Enantiomere drehen die Schwingungsebene von linear polarisiertem Licht in unterschiedlicher Orientierung. D-Alanin dreht die Ebene des polarisierten Lichts im Uhrzeigersinn und wird daher als rechtsdrehend (dextrorotatory) bezeichnet. Das L-Alanin dreht die Ebene gegen den Uhrzeigersinn und ist somit das linksdrehende (levorotatory) Enantiomer.

Um die absolute Konfiguration von Enantiomeren zu bestimmen, wurde von Cahn, Ingold und Prelog (Cahn *et al.*, 1966, Prelog *et al.*, 1982) ein System entwickelt. Hierbei wird die Umgebung um das Chiralitätszentrum betrachtet. Das Chiralitätszentrum wird in diesem Fall durch das asymmetrische Kohlenstoffatom 1002 (L-Alanin) bzw. 2002 (D-Alanin) gebildet. Dieses Atom wird von vier verschiedenen Substituenten umgeben, die unterschiedliche Prioritäten besitzen. Welche Priorität ein Substituent erhält, hängt von der Ordnungszahl der Atome ab. Die höchste Priorität erhält die Aminogruppe, da das Stickstoffatom die höchste Ordnungszahl von 7 besitzt. Die Carboxylgruppe besitzt die zweithöchste Priorität, da das Kohlenstoffatom mit Sauerstoffatomen verbunden ist, die ebenfalls eine höhere Ordnungszahl besitzen. Das Kohlenstoffatom der Methylgruppe ist hingegen nur mit Wasserstoffatomen verbunden. Die Methylgruppe erhält daher die dritthöchste Priorität. Das Wasserstoffatom besitzt mit einer Ordnungszahl von 1 die niedrigste Priorität.

Um nun die absolute Konfiguration zu bestimmen, wird das Molekül so gedreht, dass der Substituent mit der geringsten Priorität hinter die Betrachtungsebene tritt. Nimmt die Priorität der Substituenten im Uhrzeigersinn ab, besitzt das Chiralitätszentrum die absolute Konfiguration R (rectus, lateinisch, rechts). Eine S (sinister, lateinisch: links) Konfiguration des Chiralitätszentrums liegt hingegen dann vor, wenn die Priorität der Substituenten gegen den Uhrzeigersinn abnimmt.

Racemasereaktionen lassen sich mit dem Dugundji-Ugi-Modell schwer beschreiben. Bei Racemase Reaktionen kommt es weder zur Bildung oder Spaltung von Bindungen, noch kommt es zu einer Erhöhung oder Verminderung der Anzahl freier Valenzelektronen. Die Reaktionsmatrix enthält somit zunächst nur Nulleinträge. Es gibt aber die Möglichkeit die absolute Konfiguration eines Atoms entlang der Hauptdiagonalen einzutragen (Lüdge, 2002). In der vorliegenden Arbeit wurde eine S Konfiguration durch eine „-10“ und eine R Konfiguration durch eine „10“ auf der Hauptdiagonalen der Edukt- oder Produktmatrix gekennzeichnet. Ändert sich die Konfiguration eines Atoms von S nach R wird dies durch einen Matrixeintrag von 20 in der R-Matrix angezeigt (Abbildung 36). Entsprechend wird eine Konformationsänderung eines Atoms von R nach S durch einen Eintrag von „-20“ gekennzeichnet.

```
-                    C
                     1002
                     C
                     2002

C 1002   C 2002      20
```

Abbildung 36: R-Matrix der Alanin-Racemase (EC-Nummer 5.1.1.1).

2.4 Kanonisierung

Auf der Basis einer vollständigen Atomzuordnung ist es vergleichsweise einfach, die Edukt- und Produktmatrix zu generieren und die R-Matrix zu berechnen (vgl. 2.3). Schwieriger ist es hingegen, die Elektronentransfermuster der Reaktionen zu vergleichen. Die Position eines Atoms lässt sich innerhalb einer R-Matrix verschieben, solange Zeile und Spalte des Atoms den gleichen Index besitzen. Daher existieren n! mögliche Permutationen einer R-Matrix, wobei n der Dimension der R-Matrix entspricht. Um einen effizienten Vergleich zu ermöglichen, wurde deshalb ein Verfahren zur Kanonisierung von R-Matrizen verwendet.

2.4.1 Prinzipien der Kanonisierung

Die Elektronentransfermuster lassen sich nicht vergleichen, indem alle R-Matrix Permutationen erzeugt und auf Identität geprüft werden. Zum einen ist dieses Verfahren aus Laufzeit-technischer Sicht ungünstig. Beispielsweise besitzen R-Matrizen mit der Dimension 7 bereits 5040 Permutationen. Um zwei solche R-Matrizen auf Identität zu überprüfen, wären maximal 25401600 Vergleiche erforderlich. Aus diesem Grund wurden früh Kanonisierungskriterien (Brandt *et al.*, 1981, Brandt *et al.*, 1983) entwickelt, die es erlauben, aus der Vielzahl von Permutationen eine repräsentative R-Matrix zu selektieren. Die kanonische Matrix soll zwei Kriterien erfüllen.

1. kontinuierliche Folge von Einträgen $r_{ij} \neq 0$ entlang der ersten Nebendiagonalen
2. alternierende Vorzeichen der Einträge beginnend mit einem negativen Eintrag

Diese Kriterien beruhen auf der Analyse von chemischen Reaktionen. Danach bildet das Elektronentransfermuster einen Zyklus, wenn die an der Reaktion beteiligten Atome mit ihren Austauschpartnern verkettet werden. Dieses Muster resultiert aus der Beobachtung, dass ein Atom, das auf der Eduktseite an der Spaltung einer Bindung beteiligt war, mit hoher Wahrscheinlichkeit eine neue Bindung auf der Produktseite eingeht. Die Atome, mit denen es reagiert, sind wiederum ihrerseits an der Spaltung und Entstehung von anderen Bindungen beteiligt.

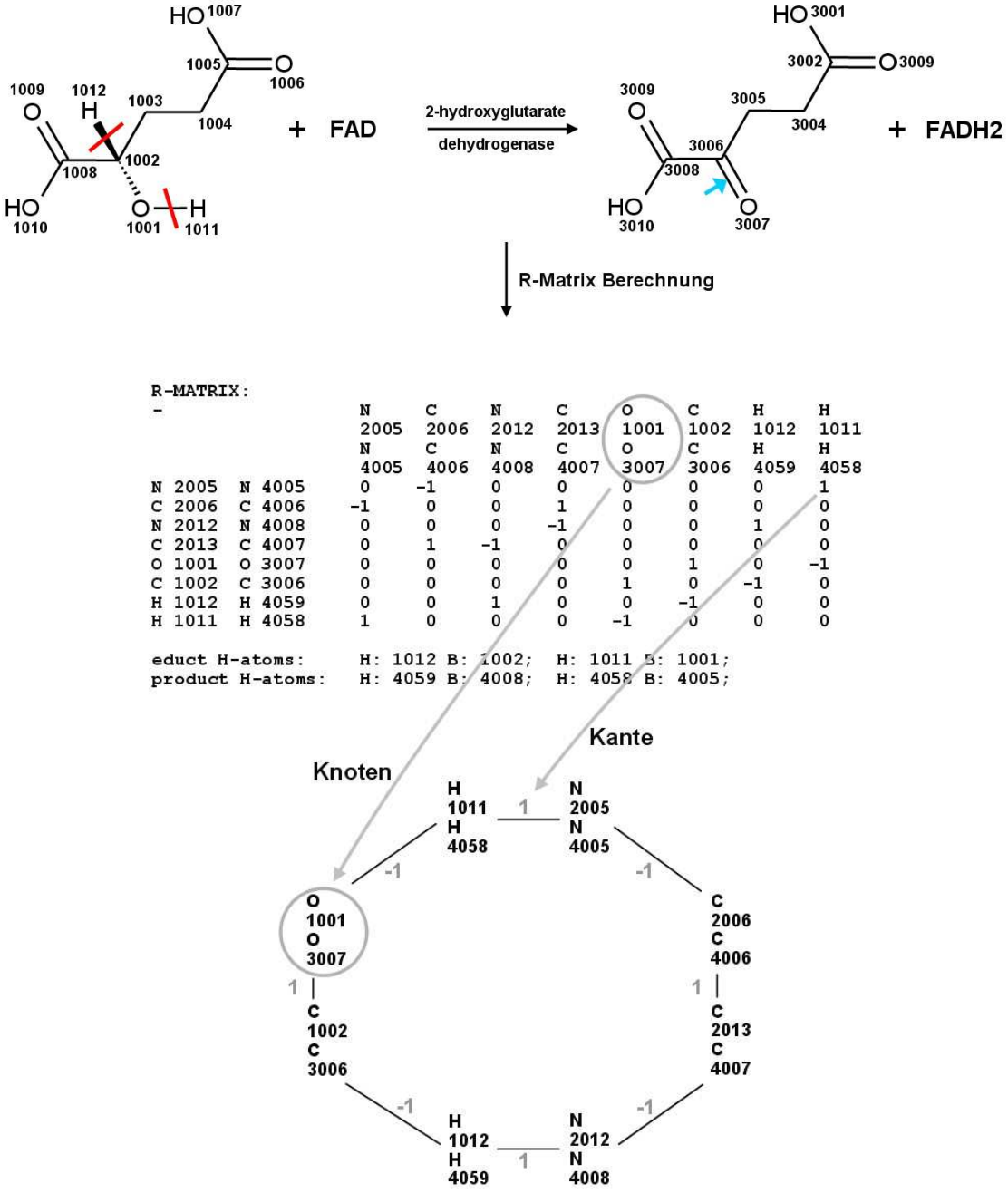


Abbildung 37: Reaktion mit zugehöriger R-Matrix der 2-Hydroxyglutarat-Dehydrogenase. Die Übertragung der R-Matrix in einen R-Graphen zeigt in welcher Wechselbeziehung die Atome der Reaktion stehen. Das Elektronentransfermuster besitzt eine zyklische Struktur.

Die Wechselbeziehungen zwischen den Atomen, die an der Reaktion beteiligt sind lassen sich anschaulich anhand eines R-Graphen darstellen.

Abbildung 37 zeigt die R-Matrix der 2-Hydroxyglutarat-Dehydrogenase. Dieses Enzym katalysiert die Oxidation von (S)-2-Hydroxyglutarat zu 2-Oxoglutarat, wobei in diesem Fall FAD reduziert

wird. Die R-Matrix hat eine Dimension von 8, womit 40320 Permutationen der R-Matrix existieren. Die R-Matrix wird in einen R-Graphen übertragen, wobei die Atome die Knoten und die Matrixeinträge die Kanten repräsentieren. Der Graph besitzt eine kreisförmige Struktur, womit der zyklische Charakter des Elektronen-Austauschmusters ersichtlich wird. Der längste Pfad innerhalb des R-Graphen gibt Aufschluss über die Reihenfolge der Atome, die für eine Kanonisierung der R-Matrix benötigt wird.

Auf der Grundlage von heuristischen Vermutungen wurden in der Vergangenheit Algorithmen zur Kanonisierung entwickelt (Brandt *et al.*, 1981, Brandt *et al.*, 1983). Diese Annahmen gehen davon aus, das Elektronentransfermuster sei zyklisch oder verliefe entlang eines einzelnen Pfades. Außerdem wird vorausgesetzt, dass sich immer ein Vorzeichenwechsel für die Einträge entlang der ersten Nebendiagonalen konstruieren lässt. Diese Algorithmen erwiesen sich jedoch für die Betrachtung biochemischer Reaktionen als unzureichend. Verzweigungen können zu unterschiedlich langen Pfaden innerhalb eines R-Graphen führen. Das Elektronen-Austauschmuster kann unterbrochen werden, wodurch mehrere Graphfragmente entstehen. In anderen Fällen ist es erforderlich, aufeinander folgende positive Einträge entlang der ersten Seitendiagonalen in Kauf zu nehmen, damit eine fortlaufende Sequenz von Einträgen erzielt wird.

Aus diesen Gründen wurde ein neuer Algorithmus zur Kanonisierung entwickelt. Wie die übrigen Module des Programms, wurde der Algorithmus in C++ implementiert und als Klasse „Canonical_Matrix“ in das Programm integriert. Er übersetzt die R-Matrix in einen R-Graphen und ermittelt mit Hilfe einer Backtracking-Suche alle längsten Pfade innerhalb des Graphen. Im Fall der 2-Hydroxyglutarat-Dehydrogenase ist der R-Graph zyklisch. Somit ist jeder der 8 Knoten ein Ausgangspunkt für zwei längste Pfade. Der Suchalgorithmus ermittelt daher zunächst 16 längste Pfade (Abbildung 38). Davon beginnen 8 Pfade mit einer negativen Kante und erfüllen somit die beiden Kanonisierungskriterien. Um die Anzahl der Pfade weiter zu reduzieren, wurde das Elementgewicht der beteiligten Atome berücksichtigt. Danach werden die Pfade aus dem Set selektiert, bei denen die schwersten Elemente eine vordere Position innerhalb des Pfades besetzen. In den meisten Fällen kann auf diese Weise ein Pfad selektiert werden. Die Reihenfolge der Atome entlang dieses Pfades ermöglicht schließlich die Konstruktion der kanonischen R-Matrix (Abbildung 39).

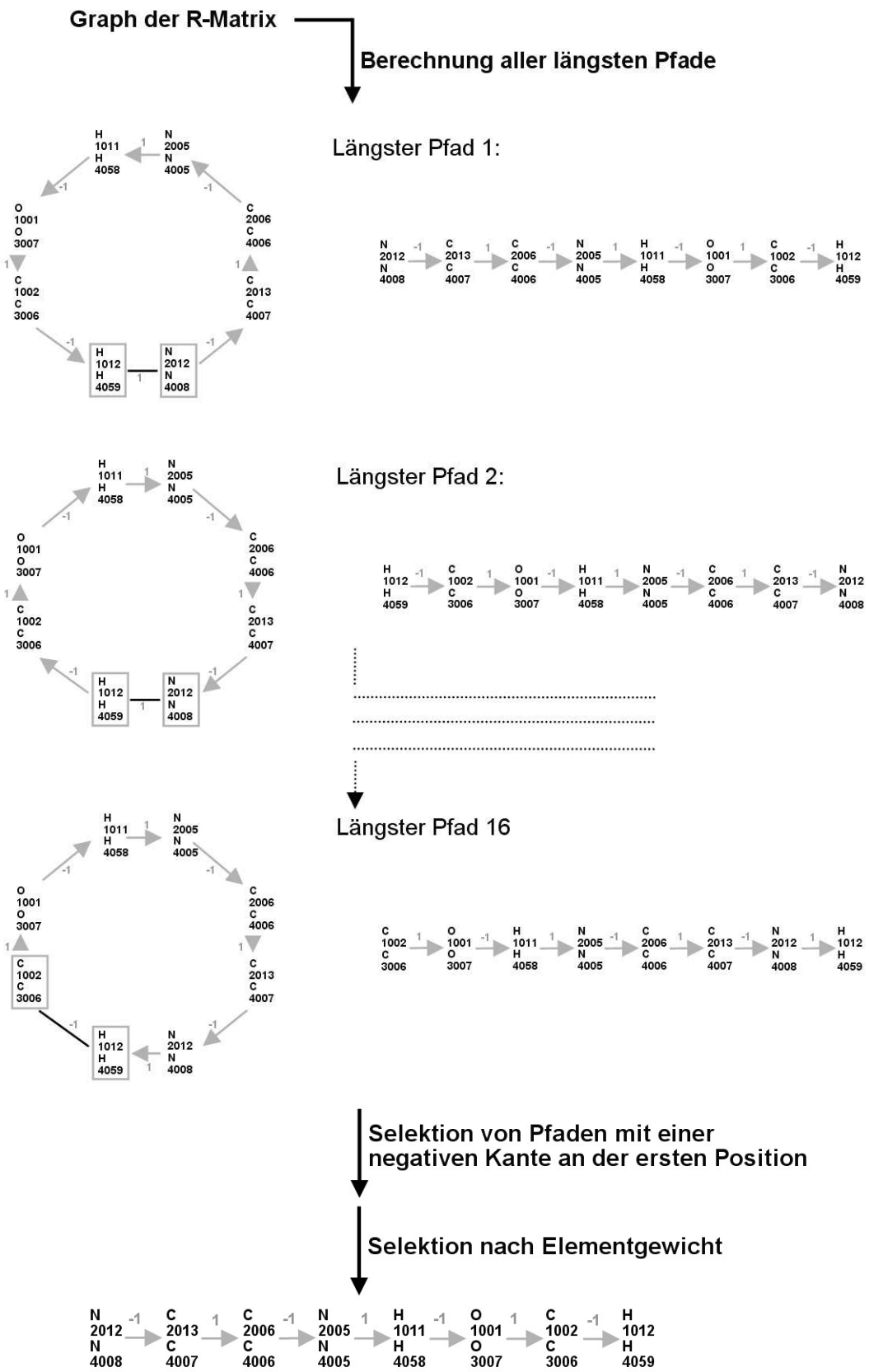
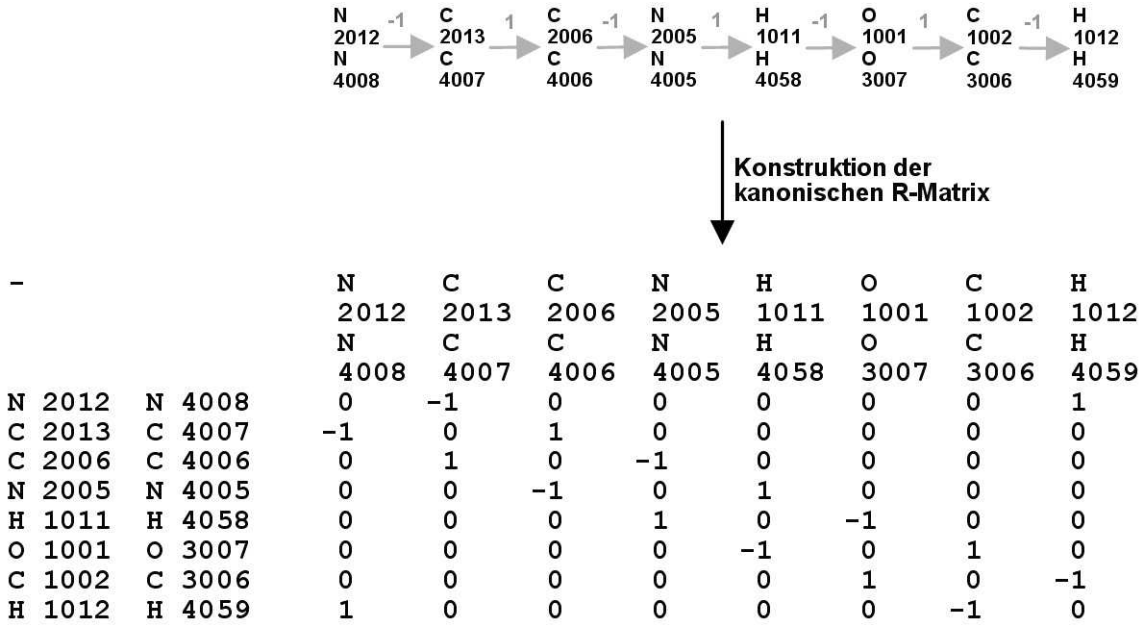


Abbildung 38: Der Backtracking-Algorithmus findet 16 längste Pfade in dem R-Graphen der R-Matrix. Durch Selektion von Pfaden mit einer Minus-Kante an erster Position und Berücksichtigung der Elementgewichte werden redundante Pfade ausgeschlossen.



-		N 2012	C 2013	C 2006	N 2005	H 1011	O 1001	C 1002	H 1012
		N 4008	C 4007	C 4006	N 4005	H 4058	O 3007	C 3006	H 4059
N 2012	N 4008	0	-1	0	0	0	0	0	1
C 2013	C 4007	-1	0	1	0	0	0	0	0
C 2006	C 4006	0	1	0	-1	0	0	0	0
N 2005	N 4005	0	0	-1	0	1	0	0	0
H 1011	H 4058	0	0	0	1	0	-1	0	0
O 1001	O 3007	0	0	0	0	-1	0	1	0
C 1002	C 3006	0	0	0	0	0	1	0	-1
H 1012	H 4059	1	0	0	0	0	0	-1	0

Abbildung 39: Mit Hilfe des längsten Pfades, der die Kanonisierungskriterien erfüllt, kann schließlich die kanonische R-Matrix konstruiert werden.

Die Reaktion der 2-Hydroxyglutarat-Dehydrogenase beschreibt einen Idealfall, für den vergleichsweise einfach eine kanonische R-Matrix errechnet werden kann. Andere Fälle sind komplexer und erfordern zusätzliche Verfahren und Selektionskriterien, die in dem folgenden Kapitel erläutert werden.

2.4.2 Kanonisierung bei fragmentierten Elektronentransfermustern

Bei einigen Reaktionen ist das Elektronentransfermuster nicht geschlossen. Häufig spielen hierbei Metallatome, Wasserstoffatome oder Veränderungen in der Stereoisomerie eine Rolle. Einige Verbindungen enthalten Eisen-, Kupfer- oder Cobaltatome, die im Reaktionsverlauf reduziert oder oxidiert werden. Die dabei aufgenommenen oder abgegebenen Elektronen führen nur zu einer Erhöhung oder Erniedrigung der Valenzelektronenzahl. Es werden daher keine Bindungen gespalten oder gebildet. Auch bei Veränderungen in der Stereoisomerie ändert sich die Konnektivität zwischen den Atomen innerhalb des Edukt- und Produktmoleküls nicht. Diese Atome sind innerhalb der R-Matrix isoliert und führen zwangläufig zu einer Fragmentierung des R-Graphen. Aber auch andere Molekülrückgratatome können den R-Graphen fragmentieren, wenn sie beispielsweise nach Verlust einer Bindung eine Ladung annehmen, anstatt eine neue Bindung einzugehen.

H^+-Ionen entstehen durch Abspaltung von Molekülen oder werden an Moleküle angebunden. Wird ein Wasserstoffatom von einem Molekül abgespalten, so ist es einerseits an der Spaltung einer Bindung beteiligt. Andererseits geht es aber auf der gegenüberliegenden Seite keine neue Bindung ein. Hierdurch entsteht ebenfalls ein Bruch in dem Elektronentransfermuster. In einigen Fällen entsteht hierdurch ein R-Graph mit linearer Form, wobei die Wasserstoffatome an den Enden des Pfades stehen. Wenn aber mehrere H^+-Ionen an einer Reaktion beteiligt sind, kommt es meist zu einer Fragmentierung des R-Graphen.

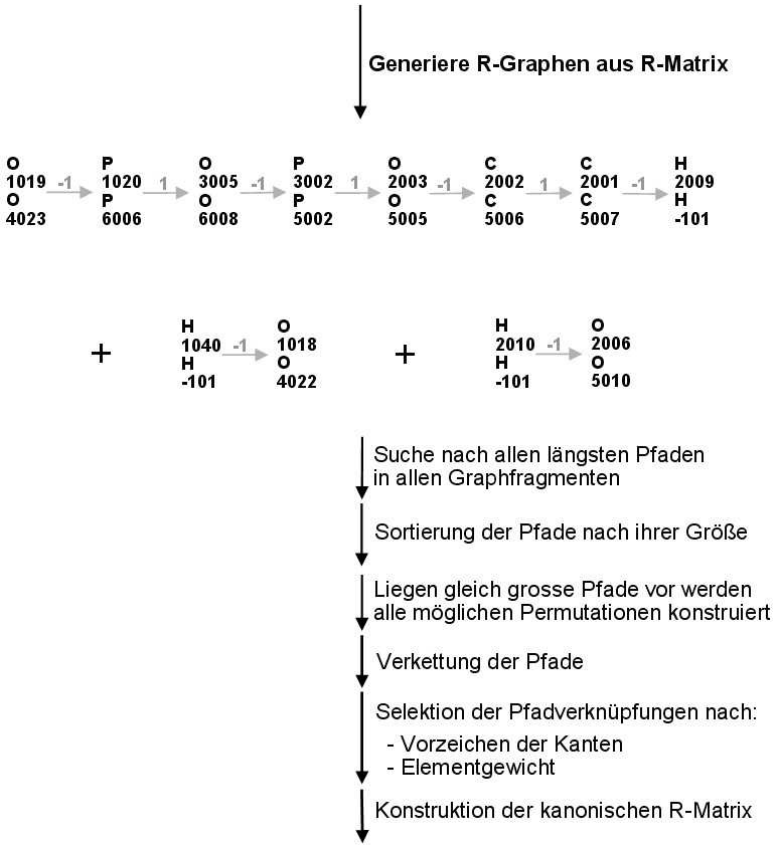
Ein Beispiel hierfür ist die Reaktion der Pyruvat-Phosphat-Dikinase (Abbildung 40). Die Übertragung der R-Matrix führt zu einem nicht zusammenhängenden Graphen, der aus 3 Fragmenten besteht. Der Backtracking-Algorithmus errechnet alle längsten Pfade innerhalb der verschiedenen Graphfragmente. Die längsten Pfade lassen sich in vielfältiger Weise verknüpfen und zu Gesamtlösungen ausbauen. Hierzu werden zunächst die Pfade nach ihrer Größe sortiert. Dabei wird den Pfaden von größeren Fragmenten Vorrang eingeräumt. Häufig kommt es vor, dass einige Pfade gleich lang sind und alle möglichen Permutationen konstruiert werden müssen. Unter Umständen können so aus einer R-Matrix einige hundert Pfadverknüpfungen entstehen. Jede Pfadverknüpfung entspricht einer Atomreihenfolge, die als Kandidat für eine kanonische R-Matrix betrachtet werden kann.

In dem folgenden Schritt wird die Anzahl der Pfadverknüpfungen reduziert. Nach Möglichkeit werden Pfadverknüpfungen selektiert, die mit einer negativen Kante beginnen. Weiterhin werden Lösungen bevorzugt, die trotz der Umbrüche dem regulären Vorzeichenwechselmuster entsprechen. Die Betrachtung des Elementgewichtes der beteiligten Atome hilft weiter, die Anzahl der Pfadverknüpfungen zu vermindern. Danach werden die Pfadverknüpfungen ausgewählt, bei denen die schwersten Elemente eine vordere Position innerhalb der Atomreihenfolge besetzen.

Im vorliegenden Beispiel wird auf diese Weise eine Pfadverknüpfung selektiert, aus der eine kanonische R-Matrix konstruiert werden kann. Entlang der ersten Seitendiagonalen wird die Folge von Einträgen zweimal durchbrochen. Das Elektronentransfermuster ist nicht zusammenhängend, wie dies bereits aus dem R-Graphen zu ersehen war.

Anhand der Beispiele wird das allgemeine Prinzip der Kanonisierung ersichtlich. Ein anderer wichtiger Aspekt bei der Kanonisierung ist die Zuordnung der Wasserstoffatome, die an der Reaktion beteiligt sind. Wasserstoffatome, die auf der Eduktseite abgespalten werden und auf der Produktseite eine neue Bindung eingehen, unterbrechen nicht das Elektronentransfermuster. Anders ist dies bei Wasserstoffatomen, die als H^+-Ionen in eine Reaktion eingehen oder aus ihr hervor-

-		C 2001 C 5007	C 2002 C 5006	O 2006 O 5010	O 2003 O 5005	P 3002 P 5002	O 1018 O 4023	O 1019 O 4022	P 1020 P 6006	O 3004 O 6008	H 1040 H -101	H 2010 H -101	H 2009 H -101
C 2001	C 5007	0	1	0	0	0	0	0	0	0	0	0	-1
C 2002	C 5006	1	0	0	-1	0	0	0	0	0	0	0	0
O 2006	O 5010	0	0	0	0	0	0	0	0	0	0	-1	0
O 2003	O 5005	0	-1	0	0	1	0	0	0	0	0	0	0
P 3002	P 5002	0	0	0	1	0	0	0	0	-1	0	0	0
O 1018	O 4023	0	0	0	0	0	0	0	0	0	-1	0	0
O 1019	O 4022	0	0	0	0	0	0	0	-1	0	0	0	0
P 1020	P 6006	0	0	0	0	0	0	-1	0	1	0	0	0
O 3004	O 6008	0	0	0	0	-1	0	0	1	0	0	0	0
H 1040	H -101	0	0	0	0	0	-1	0	0	0	0	0	0
H 2010	H -101	0	0	-1	0	0	0	0	0	0	0	0	0
H 2009	H -101	-1	0	0	0	0	0	0	0	0	0	0	0



-		O 1019 O 4023	P 1020 P 6006	O 3005 O 6008	P 3002 P 5002	O 2003 O 5005	C 2002 C 5006	C 2001 C 5007	H 2009 H -101	O 1018 O 4022	H 1040 H -101	O 2006 O 5010	H 2010 H -101
O 1019	O 4023	0	-1	0	0	0	0	0	0	0	0	0	0
P 1020	P 6006	-1	0	1	0	0	0	0	0	0	0	0	0
O 3005	O 6008	0	1	0	-1	0	0	0	0	0	0	0	0
P 3002	P 5002	0	0	-1	0	1	0	0	0	0	0	0	0
O 2003	O 5005	0	0	0	1	0	-1	0	0	0	0	0	0
C 2002	C 5006	0	0	0	0	-1	0	1	0	0	0	0	0
C 2001	C 5007	0	0	0	0	0	1	0	-1	0	0	0	0
H 2009	H -101	0	0	0	0	0	0	-1	0	0	0	0	0
O 1018	O 4022	0	0	0	0	0	0	0	0	-1	0	0	0
H 1040	H -101	0	0	0	0	0	0	0	-1	0	0	0	0
O 2006	O 5010	0	0	0	0	0	0	0	0	0	0	-1	0
H 2010	H -101	0	0	0	0	0	0	0	0	0	-1	0	0

Abbildung 40: Aus der R-Matrix der Reaktion der Pyruvat-Phosphat-Dikinase (EC-Nummer 2.7.9.1) wird systematisch die entsprechende kanonische R-Matrix konstruiert.

gehen. In einigen Fällen lassen sich Umbrüche im R-Graphen vermeiden, wenn die Zuordnung der Wasserstoffatome vertauscht wird. Hierbei kann das H^+-Ion, das den Umbruch verursacht, an das Ende des Elektronentransfermusters verschoben werden. Die beiden Graphfragmente hingegen werden durch das Wasserstoffatom verbunden, welches mit dem H^+-Ion vertauscht wurde.
Um möglichst verbundene Elektronentransfermuster zu erhalten, müssen somit alle möglichen Wasserstoffatomzuordnungen konstruiert und die zugehörigen R-Matrizen errechnet werden. Anschließend werden die R-Matrizen selektiert, die zu möglichst zusammenhängenden R-Graphen führen. Die Kanonisierung folgt somit dem in Abbildung 41 beschriebenen Verfahren.

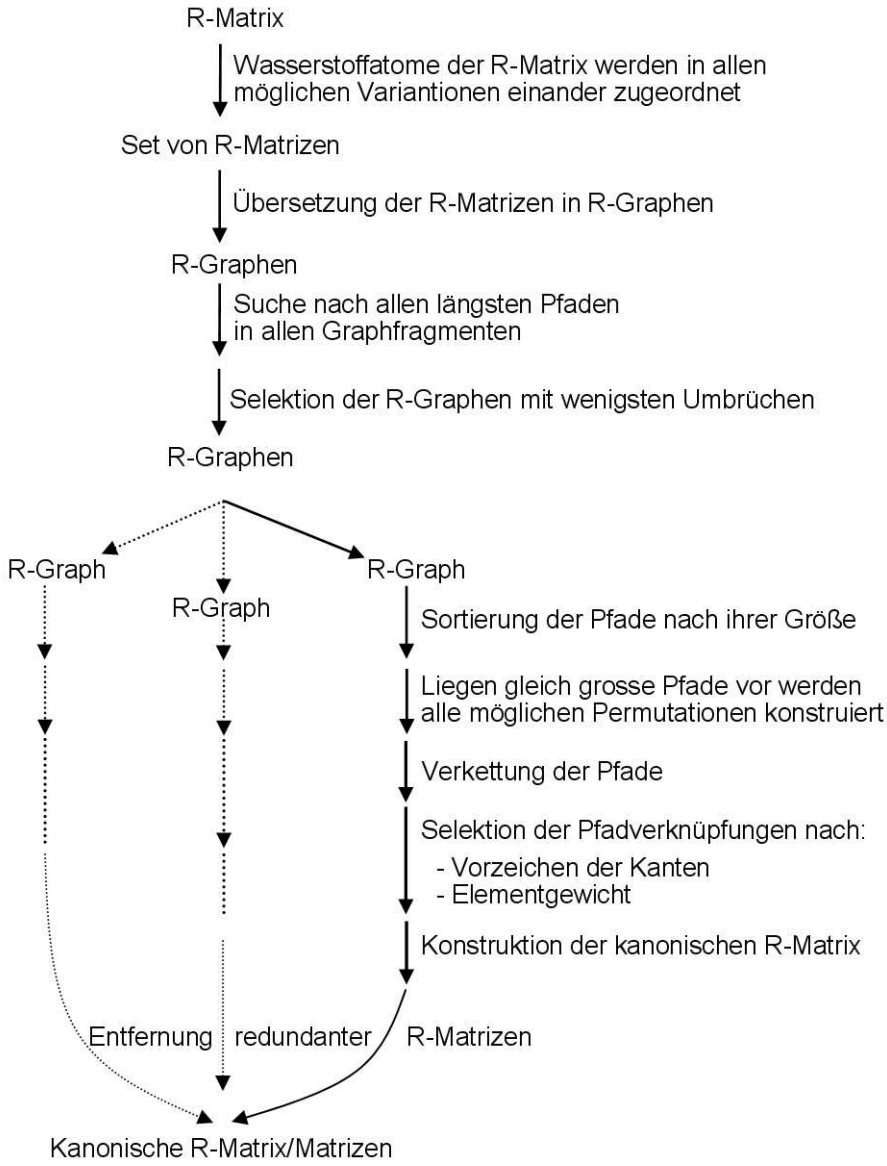


Abbildung 41: Kanonisierungsschema

2.4.3 R-Strings

In den meisten Fällen erlaubt es die Kanonisierung eine eindeutige R-Matrix zu konstruieren. Nur bei wenigen Reaktionen werden 2 oder 4 R-Matrizen erzeugt, die den Kanonisierungskriterien entsprechen. Um den Umgang mit den Elektronentransfermustern weiter zu vereinfachen wird die kanonisierte R-Matrix in eine String-Notation überführt. Strings lassen sich einfacher vergleichen und erleichtern die Gruppierung der Reaktionen anhand ihrer Elektronentransfermuster. Zur Erzeugung dieser R-Strings werden die Matrixelemente nach einem definiertem System verkettet. Ausgangspunkt ist der Matrixeintrag an der ersten Position der ersten Seitendiagonalen. An diesen werden die übrigen Einträge entlang der ersten Seitendiagonalen angehängt. Die erste Seitendiagonale wird um das Matrixelement $r_{n,1}$ erweitert, denn ein Eintrag $r_{n,1} \neq 0$ bedeutet, dass das letzte Atom in der Matrix mit dem zweiten und vorletzten Atom über einen Matrixeintrag in Verbindung steht. Demnach führt die Verkettung der Elemente der ersten Seitendiagonalen in dem Beispiel von Abbildung 42 zu dem Substring „ABCF“. In ähnlicher Weise werden auch die anderen Seitendiagonalen verlängert, um den Verlauf des Elektronentransfermusters nachzuvollziehen.

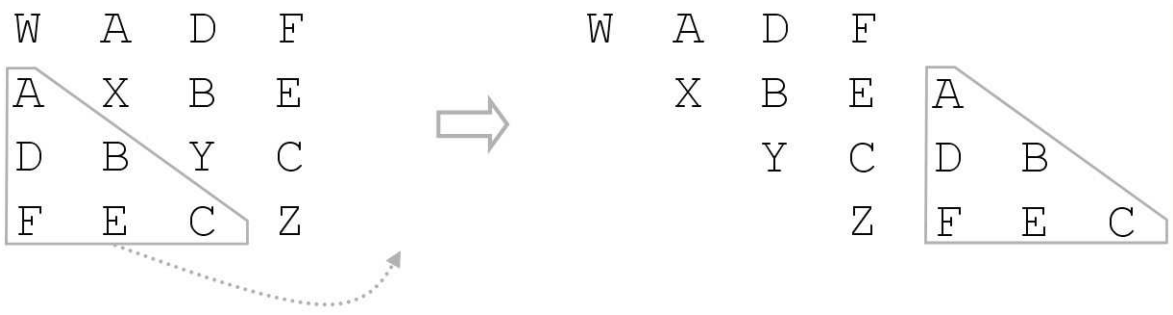


Abbildung 42: Zur Erzeugung des R-Strings werden die Einträge der Seitendiagonalen so aus der R-Matrix ausgelesen, als ob sich die untere Matrixhälfte an der rechten Seite der Matrix befände.

Durch Verkettung der Substrings der Seitendiagonalen wird der String „ABCFDEDEFABC“ generiert. Da die Matrix symmetrisch ist, lässt sich der String ohne Verlust von Information auf den Abschnitt „ABCFDE“ verkürzen.

Entlang der Hauptdiagonalen stehen die Veränderungen in der Anzahl der freien Valenzelektronen eines Atoms. Die Einträge der Hauptdiagonalen werden ebenfalls verkettet und zu dem String der Seitendiagonalen addiert. Hierdurch entsteht ein Integer-String mit den Elementen „ABCFDEWXYZ“.

Neben dem Integer-String wird ein String aus Char-Werten erzeugt, in dem die Elementsymbole der beteiligten Atome abgespeichert werden. Der Char-String bezieht sich auf die Einträge des Integer-

String. Für die Seitendiagonalelemente der R-Matrix werden 2 Elementsymbole in den Char-String aufgenommen, da je 2 Atome an der Spaltung oder Entstehung von Bindungen beteiligt sind. Die Einträge auf der Hauptdiagonalen der R-Matrix beziehen sich hingegen auf ein Atom, womit nur ein Elementsymbol des dem Char-String hinzugefügt werden darf.

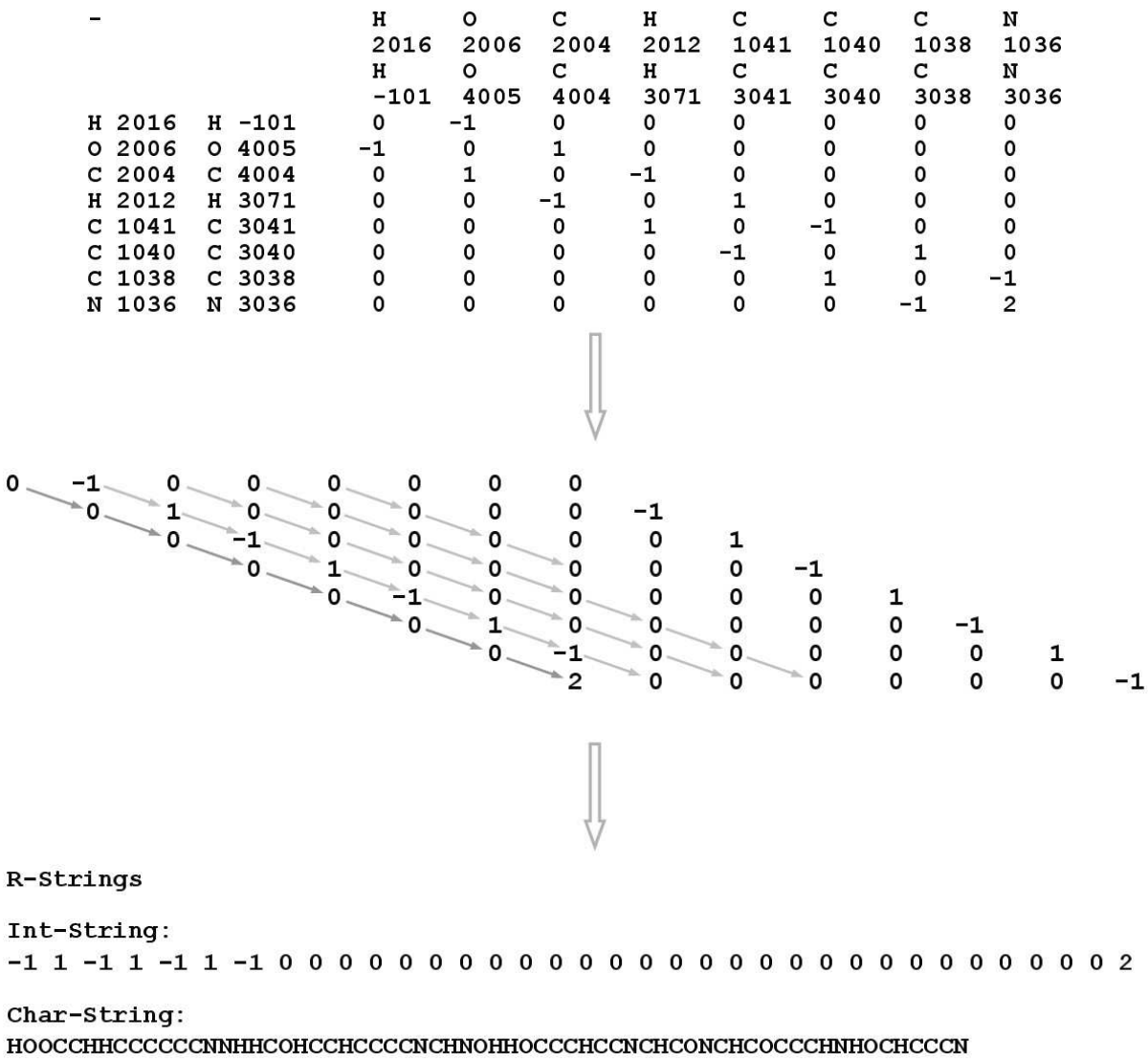


Abbildung: 43: Erzeugung der R-Strings für die Reaktion der (R,R)-Butanediol-Dehydrogenase (EC-Nummer 1.1.1.4). Die Einträge entlang der Diagonalen werden verkettet und zu einem Integer-String umgesetzt. Um zusätzlich den Atomtyp zu berücksichtigen, wird ein Char-String aus den Elementsymbolen der beteiligten Atome generiert.

3. Ergebnisse

Das automatisierte Verfahren zur R-Matrix-Berechnung ermöglicht die Analyse eines großen Sets von enzymatischen Reaktionen. Mit Hilfe der Kanonisierung und der Konvertierung zu R-Strings lassen sich Reaktionen einfach charakterisieren und vergleichen. Ein Schwerpunkt dieser Arbeit lag daher sowohl in dem Vergleich als auch in einer neuen Gruppierung der Enzyme anhand ihres Elektronentransfermusters. Um einen Vergleich der neuen Gruppen gegenüber dem EC-Klassifikationssystem zu ermöglichen, wurden die kanonischen R-Matrizen der neuen Gruppen charakterisiert und die durch das EC-System vorgegebenen Subsubklassen auf Homogenität in ihren Elektronentransfermustern untersucht.

3.1 Datenset

Die Grundlage für die Berechnung der R-Matrizen und R-Strings bildete ein Set von 3330 enzymatischen Reaktionen. Zu dem Set gehören 3876 Moleküle, die in einem definierten Format (MDL-Format) in Textdateien, den so genannten „Molfiles“, abgespeichert sind. Das Programm zur R-Matrix-Berechnung benötigt nur diese Textdateien als Eingabe zusammen mit der Information, welche Moleküle die Edukte oder Produkte darstellen. Das Datenset stammt aus der Braunschweiger Enzymdatenbank BRENDA (Schomburg, 2004). Es wurde in den letzten Jahren kontinuierlich erweitert und deckt 228 der bislang formulierten Subsubklassen ab. Die Moleküle wurden so konstruiert, dass durch die Reaktion bedingte Veränderungen automatisch erkannt werden können.
In der ersten Phase der vorliegenden Arbeit wurden Molfiles der KEGG-Datenbank (Kyoto Encyclopedia of Genes and Genomes) verwendet. Die KEGG-Datenbank (Kanehisa, 2008) weist ebenfalls ein umfangreiches Set enzymatischer Reaktionen auf. Allerdings werden viele Subsubklassen nicht abgedeckt. Besonders bei Reaktionen, wo langkettige Polymere beteiligt sind, finden sich keine Molfiles oder die Moleküle werden auf der Edukt- und Produktseite durch identische Moleküle beschrieben.

3.2 Matrixdimension der Subsubklassen

Die Subsubklassen werden in Tabelle 4 nach den Dimensionen ihrer R-Matrizen eingeteilt. Für jede Subsubklasse wird die Dimension der häufigsten R-Matrix aufgeführt. Die R-Matrixdimension entspricht der Anzahl der Atome, die an der Reaktion beteiligt sind. Bei 190 Subsubklassen und somit bei dem überwiegenden Teil der Reaktionen entspricht die Matrixdimension einer geraden Zahl. Die geraden R-Matrixdimensionen resultieren aus der Begebenheit, dass bei der Entstehung oder Spaltung einer Bindung je 2 Atome beteiligt sind. Ungerade R-Matrixdimensionen ergeben

Matrixdimension	Subsubklassen					
1	5.1.1.-	5.1.2.-	5.1.3.-	5.1.99.-		
2	1.9.99.-					
3	1.12.99.6-					
4	1.1.99.-	2.3.2.-	2.7.13.-	3.1.16.-	3.4.19.-	3.10.1.-
	1.2.99.-	2.4.1.-	2.7.99.-	3.1.21.-	3.4.21.-	3.11.1.-
	1.3.99.-	2.4.2.-	2.8.2.-	3.1.25.-	3.4.22.-	4.1.1.-
	1.7.99.-	2.4.99.-	2.8.3.-	3.1.26.-	3.4.23.-	4.1.2.-
	1.11.1.-	2.6.3.-	2.8.4.-	3.1.30.-	3.4.24.-	4.1.3.-
	1.12.2.-	2.6.99.-	3.1.1.-	3.2.1.-	3.4.25.-	4.2.1.-
	1.12.7.-	2.7.1.-	3.1.2.-	3.2.2.-	3.5.1.-	4.2.99.-
	1.13.11.-	2.7.2.-	3.1.3.-	3.3.1.-	3.5.2.-	4.3.1.-
	1.17.99.-	2.7.3.-	3.1.4.-	3.3.2.-	3.6.1.-	4.3.2.-
	1.20.98.-	2.7.4.-	3.1.5.-	3.4.11.-	3.6.2.-	4.3.3.-
	1.20.99.-	2.7.6.-	3.1.6.-	3.4.13.-	3.6.3.-	4.5.1.-
	1.97.1.-	2.7.7.-	3.1.7.-	3.4.14.-	3.6.4.-	4.6.1.-
	2.1.1.-	2.7.8.-	3.1.8.-	3.4.15.-	3.6.5.-	4.99.1.-
	2.1.3.-	2.7.10.-	3.1.11.-	3.4.16.-	3.7.1.-	
	2.1.4.-	2.7.11.-	3.1.13.-	3.4.17.-	3.8.1.-	
	2.3.1.-	2.7.12.-	3.1.15.-	3.4.18.-	3.9.1.-	
5	1.6.99.-	1.15.1.-	5.2.1.-	5.4.1.-	5.99.1.-	
	1.8.99.-	2.6.1.-	5.3.2.-	5.4.4.-		
	1.12.1.-	3.13.1.-	5.3.3.-	5.5.1.-		
6	1.1.2.-	1.3.3.-	1.14.99.-	2.2.1.-	3.12.1.-	6.2.1.-
	1.1.3.-	1.4.4.-	1.16.8.-	2.8.1.-	4.2.2.-	6.3.1.-
	1.2.2.-	1.4.99.-	1.17.4.-	2.9.1.-	4.2.3.-	6.3.2.-
	1.2.3.-	1.5.99.-	1.20.4.-	3.5.3.-	5.4.3.-	6.3.3.-
	1.2.4.-	1.7.2.-	1.21.99.-	3.5.4.-	5.4.99.-	6.4.1.-
	1.3.2.-	1.12.98.-	2.1.2.-	3.5.99.-	6.1.1.-	
7	1.1.4.-	1.8.3.-	1.13.99.-	1.18.1.-	3.1.27.-	5.3.1.-
	1.6.2.-	1.8.4.-	1.14.18.-	3.1.14.-	3.1.31.-	
	1.8.2.-	1.13.12.-	1.16.1.-	3.1.22.-	3.5.5.-	
8	1.1.1.-	1.5.3.-	1.8.5.-	1.14.15.-	1.20.1.-	4.4.1.-
	1.2.1.-	1.6.3.-	1.8.98.-	1.14.16.-	2.3.3.-	6.3.4.-
	1.3.1.-	1.6.6.-	1.10.99.-	1.14.19.-	2.5.1.-	6.3.5.-
	1.4.2.-	1.7.3.-	1.12.5.-	1.17.1.-	2.7.9.-	
	1.4.3.-	1.8.1.-	1.14.11.-	1.17.3.-	4.1.99.-	
9	1.5.4.-	1.6.1.-	1.9.6.-	5.3.99.-	6.5.1.-	
10	1.1.5.-	1.4.1.-	1.5.8.-	1.14.13.-	1.16.3.-	
	1.2.7.-	1.4.7.-	1.9.3.-	1.14.14.-	1.21.4.-	
	1.3.5.-	1.5.1.-	1.14.12.-	1.14.17.-	5.4.2.-	
11	1.10.2.-					
12	1.3.7.-	1.5.5.-	1.6.5.-	1.14.20.-	1.14.21.-	
15	1.7.1.-					
16	1.5.7.-	1.10.1.-				
18	1.17.5.-	1.21.3.-				
21	1.7.7.-					
22	1.8.7.-	1.10.3.-	6.6.1.-			
24	1.19.6.-					
90	1.18.6.-					

Tabelle 4: Gliederung der Subsubklassen nach der Matrixdimension ihrer häufigsten R-Matrix.

einer Bindung je 2 Atome beteiligt sind. Ungerade R-Matrixdimensionen ergeben sich, wenn ein Atom an der Entstehung bzw. Spaltung von mehreren Bindungen beteiligt ist oder sich eine Veränderung nur auf ein Atom auswirkt (Valenzelektronenzahl, stereochemische Veränderungen).
93 Subsubklassen weisen eine R-Matrixdimension von 4 auf. Bei den meisten Reaktionen mit dieser Dimension werden 2 Bindungen auf der Eduktseite gespalten, während auf der Produktseite 2 Bindungen entstehen. Dieser Reaktionstyp ist unter den Enzymen am weitesten verbreitet. Die niedrigste R-Matrixdimension von Eins besitzen die Reaktionen der Subklasse 5.1.-.- Bei den Reaktionen dieser Isomerasen ändert sich nur die Stereoisomerie eines Atoms. Die größte Matrixdimension von 90 besitzt die Reaktion der Nitrogenase (EC-Nummer 1.18.6.1). Bei dieser Reaktion reagieren 41 Eduktmoleküle und 8 H^+ Atome zu 42 Produktmolekülen. An der Reaktion sind überwiegend Kofaktoren (ATP und Ferredoxin) beteiligt.
An den Reaktionen der ersten Hauptklasse sind meistens Kofaktoren in geringer oder hoher Anzahl beteiligt, die als Elektronendonatoren oder Akzeptoren fungieren. Kofaktoren haben bedeutenden Einfluss auf die Matrixdimensionen. Beispielsweise erhöht die Umwandlung von NAD^+ zu NADH die Matrixdimension um 5 Zeilen und Spalten, während die Beteiligung eines Cytochrom-Moleküls zu einer Erweiterung der R-Matrix um eine Zeile und Spalte führt. Die Oxidoreduktasen sind daher fast in allen vertretenen R-Matrixdimensionen zu finden. Ab einer R-Matrixdimension von 10 sind ausschließlich nur noch Oxidoreduktasen vertreten.
Die Transferasen, Hydrolasen und Lyasen besitzen überwiegend eine Matrixdimension von 4. Bei den Transfereasereaktionen wird eine Molekülstruktur von einem Molekül auf ein anderes Molekül übertragen. Hierzu ist meist auf der Eduktseite die Spaltung von 2 Bindungen erforderlich, während auf der Produktseite 2 neue Bindungen entstehen. Auch die Spaltung von Molekülen folgt im Wesentlichen diesem Reaktionsmuster, ob nun mit der Beteiligung eines Wassermoleküls, wie bei den Hydrolasen, oder ohne die Beteiligung eines Wassermoleküls, wie bei den Lyasen.

3.3 Homogenität der Subsubklassen

Für die Subsubklassen war eine hohe Homogenität der Elektronentransfermuster zu erwarten. Die Reaktionen sind innerhalb der Subsubklassen bereits so weit spezifiziert, dass die beteiligten Bindungen, Elektronenakzeptoren, Kofaktoren und stereochemischen Veränderungen meist identisch sind. Tatsächlich konnte eine durchschnittliche Übereinstimmung der kanonischen R-Matrizen von 76 Prozent innerhalb der Subsubklassen errechnet werden. Das Kreisdiagramm in Abbildung 44 gibt Aufschluss über die Verteilung der R-Stringidentität unter den Subsubklassen. Mehr als 59 Prozent der Subsubklassen haben jeweils ein fast identisches Elektronentransfermuster. Die Identität der R-

Strings liegt zwischen 75 und 100 Prozent. Bei 23 Prozent der Subsubklassen liegt die Identität immer noch zwischen 50 und 75 Prozent. 18 Prozent der Subsubklassen besitzen hingegen eine sehr geringe R-Stringidentität unterhalb von 50 Prozent.

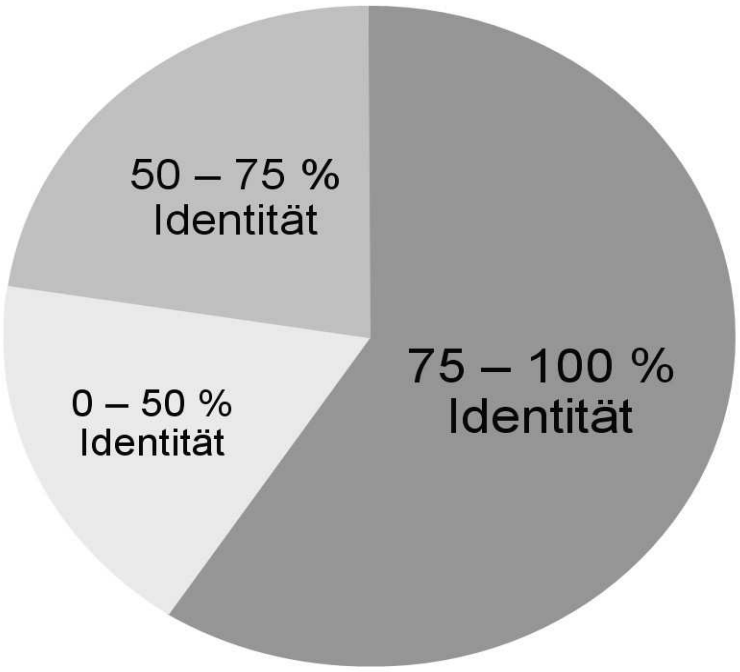


Abbildung 44: Verteilung der R-Stringidentität unter den Subsubklassen des EC-Klassifikationssystems.

Bei der Analyse der R-Stringidentität muss die sehr unterschiedliche Zusammensetzung der Subsubklassen berücksichtigt werden. Bei einigen Enzymgruppen sind nur wenige Vertreter eines Reaktionstyps bekannt oder die katalysierte Reaktion lässt sich nur anhand einer allgemeinen Formel beschreiben. Sechzig Subsubklassen des Datensets werden nur durch die Reaktion einer EC-Nummer vertreten (vgl. Tabelle 5). Wird die Anzahl der EC-Nummern jeder Subsubklasse entsprechend gewichtet, ergibt sich eine durchschnittliche R-Stringidentität von 69,45 Prozent.

Die sehr geringe Homogenität einiger Subsubklassen in Hinblick auf ihre Elektronentransfermuster kann verschiedene Ursachen haben. Beispielsweise können Parallelreaktionen durch weitere an der Reaktion beteiligte Moleküle auftreten. Es können unterschiedliche Kofaktoren als Elektronendonatoren oder Akzeptoren fungieren, oder verschiedene funktionelle Gruppen übertragen werden.

Der Anteil von Subsubklassen mit geringer Homogenität enthält einige Subsubklassen des Typs a.b.98.- oder a.b.99.-. Diese Subsubklassen bestehen aus enzymatischen Reaktionen, die zwar eine gemeinsame Eigenschaft teilen, aber zum Teil verschiedene Akzeptoren benutzen, andere Gruppen übertragen oder sonstige Unterschiede aufweisen. Ein Beispiel hierfür ist die Subsubklasse 1.10.99.-,

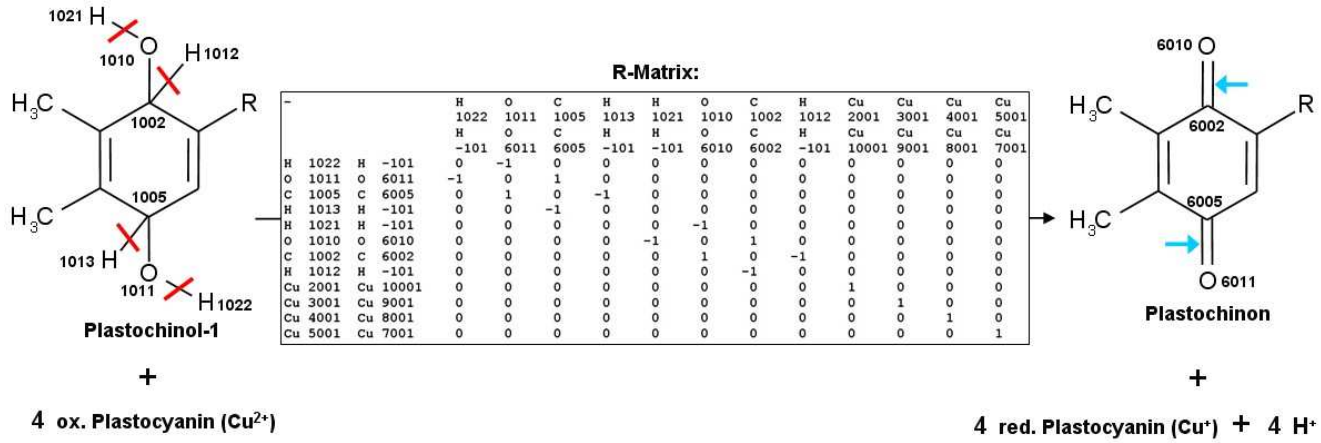


Abbildung 45: Die kanonische R-Matrix der Plastochinol-Plastocyanin-Reduktase (EC-Nummer 1.10.99.1) ist auch innerhalb der Subsubklasse 1.10.99.- einmalig. In keiner anderen Reaktion der Subsubklasse ist Plastocyanin beteiligt oder werden Cu^{2+}-Atome reduziert.

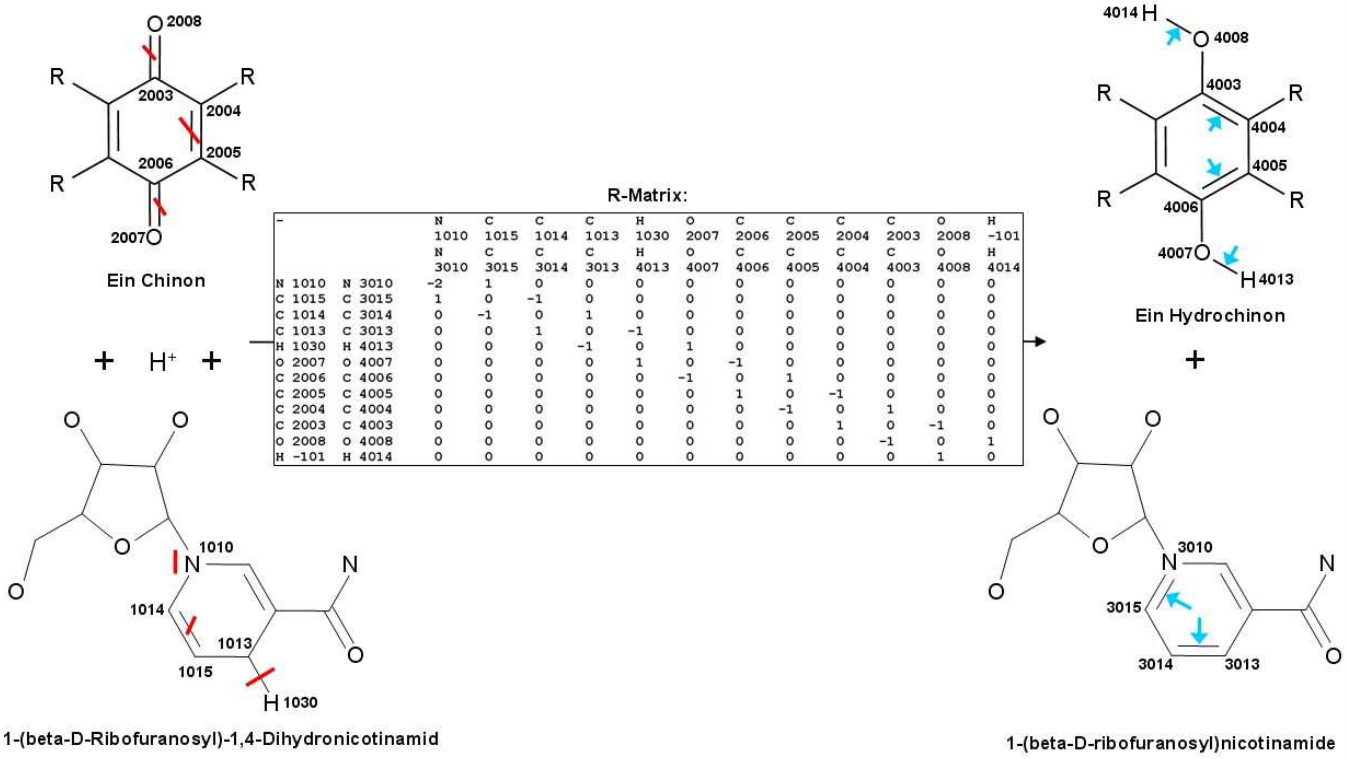
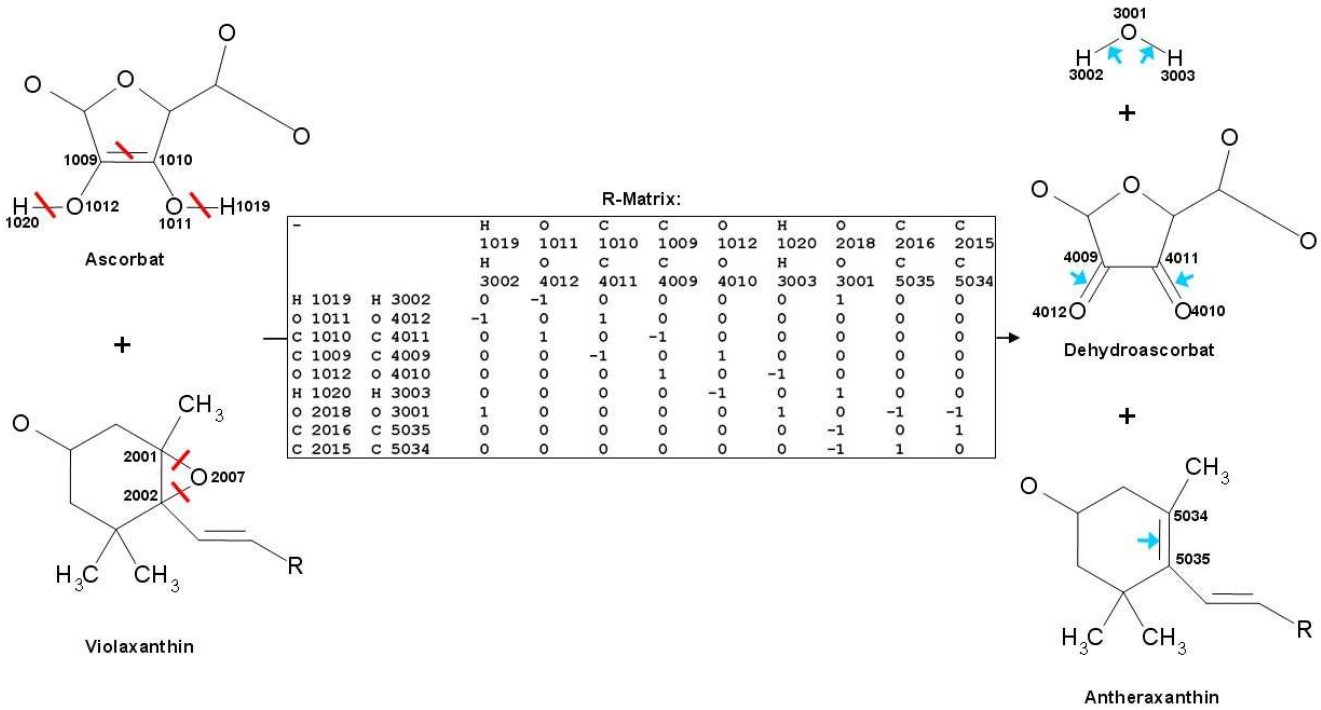


Abbildung 46: Kanonische R-Matrix der Reduktaseribosyldihydronicotinamid-Dehydrogenase (EC-Nummer 1.10.99.2). Die Oxidation eines Chinons beinhaltet andere Bindungsumbrüche als die Oxidation von Plastochinol (EC-1.10.99.1) oder Ascorbat (EC-1.10.99.3). Die R-Matrix ist reaktionsspezifisch.



-		H 1019 H 3002	O 1011 O 4012	C 1010 C 4011	C 1009 C 4009	O 1012 O 4010	H 1020 H 3003	O 2018 O 3001	C 2016 C 5035	C 2015 C 5034
H 1019	H 3002	0	-1	0	0	0	0	1	0	0
O 1011	O 4012	-1	0	1	0	0	0	0	0	0
C 1010	C 4011	0	1	0	-1	0	0	0	0	0
C 1009	C 4009	0	0	-1	0	1	0	0	0	0
O 1012	O 4010	0	0	0	1	0	-1	0	0	0
H 1020	H 3003	0	0	0	0	-1	0	1	0	0
O 2018	O 3001	1	0	0	0	0	1	0	-1	-1
C 2016	C 5035	0	0	0	0	0	0	-1	0	1
C 2015	C 5034	0	0	0	0	0	0	-1	1	0

Abbildung 47: Kanonische R-Matrix der Violaxanthin de-Epoxidase (EC-Nummer 1.10.99.3). Beide Eduktmoleküle der Violaxanthin de-Epoxidase werden oxidiert. Dabei wird ein Sauerstoffatom des Violaxanthin abgespalten, das zu einem Wassermolekül reduziert wird.

die drei enzymatische Reaktionen enthält. Die Enzyme dieser Subsubklasse gehören zu den Oxidoreduktasen und benutzen als gemeinsames Merkmal Diphenole oder verwandte Substanzen als Elektronendonatoren. Wie in den Abbildungen 45 bis 47 dargestellt unterscheiden sich alle Reaktionen der Subsubklasse.

Als Donatoren und Akzeptoren dienen jeweils unterschiedliche Moleküle. Dies führt zu verschiedenen strukturellen Veränderungen und Elektronentransfermustern. Es ergibt sich für die Subsubklasse nur eine Matrixidentität von 33,33 Prozent. Sie enthält ein Set von Reaktionen, die in keine andere Subsubklasse der Subklasse eingeordnet werden konnten. Ebenso erklärt sich die geringe Homogenität der Subsubklassen 1.2.99.-, 1,.3.99.-, 1.5.99.-, 1.7.99.-, 1.12.98.-, 1.13.99.-, 1.14.99.-, 1.17.99.-.

Neben den Subsubklassen der Form a.b.98.- oder a.b.99.- finden sich noch weitere Subsubklassen mit geringer Homogenität in den Elektronentransfermustern. Die Gründe hierfür sind vielfältig. Häufig gibt es innerhalb einer Subsubklasse größere und kleinere Gruppen mit identischer kanonischer R-Matrix. Ein Beispiel hierfür ist die Subsubklasse 1.8.4.-, die nur eine geringe R-Stringidentität von 27,27 Prozent aufweist. Das Datenset dieser Subsubklasse enthält 11 Reaktionen, die 6 Gruppen mit identischem R-String bilden (Abbildung 48 bis 53). Bei fast allen Reaktionen

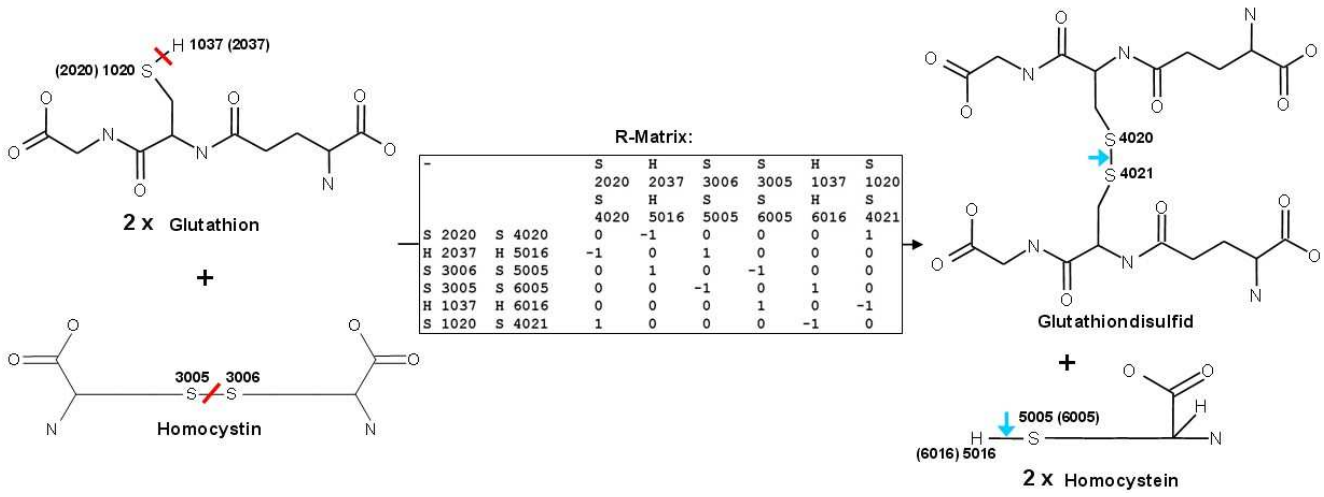


Abbildung 48: Reaktion der Glutathion-Homocystin-Transhydrogenase (EC-Nummer 1.8.4.1). Die kanonische R-Matrix repräsentiert auch die Reaktion der Glutathion-Cystin-Transhydrogenase (EC-Nummer 1.8.4.4).

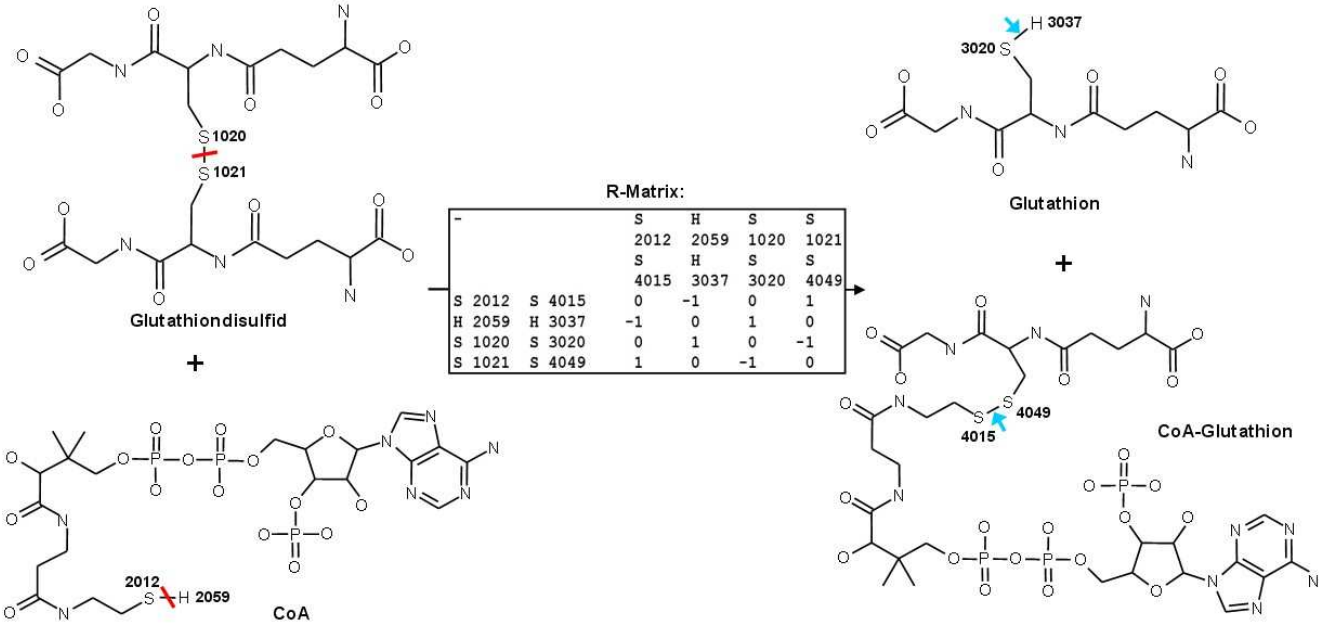
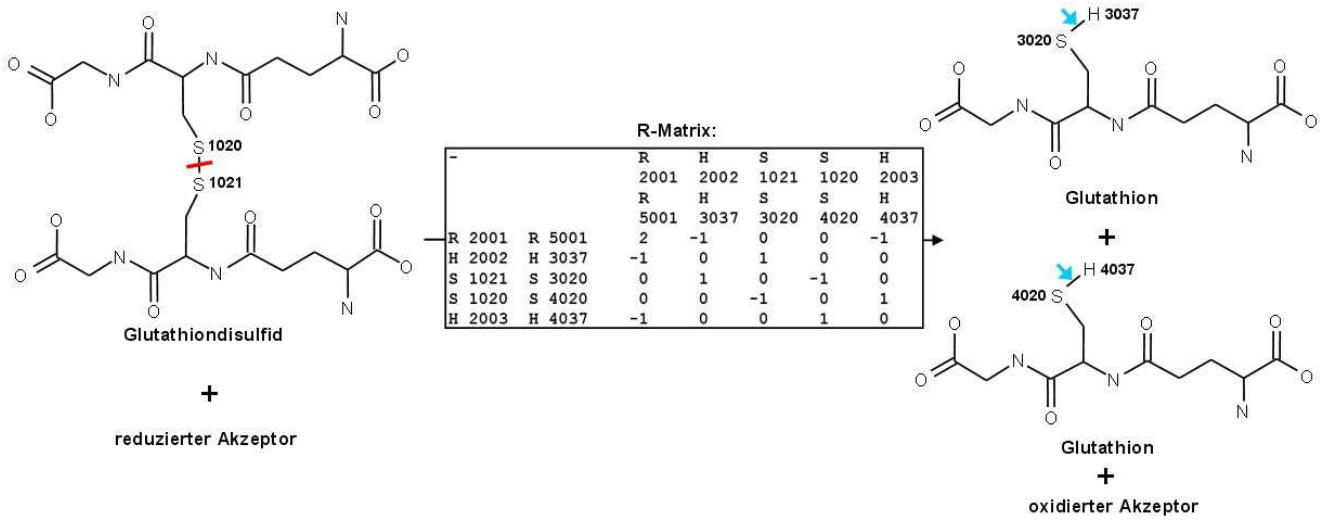
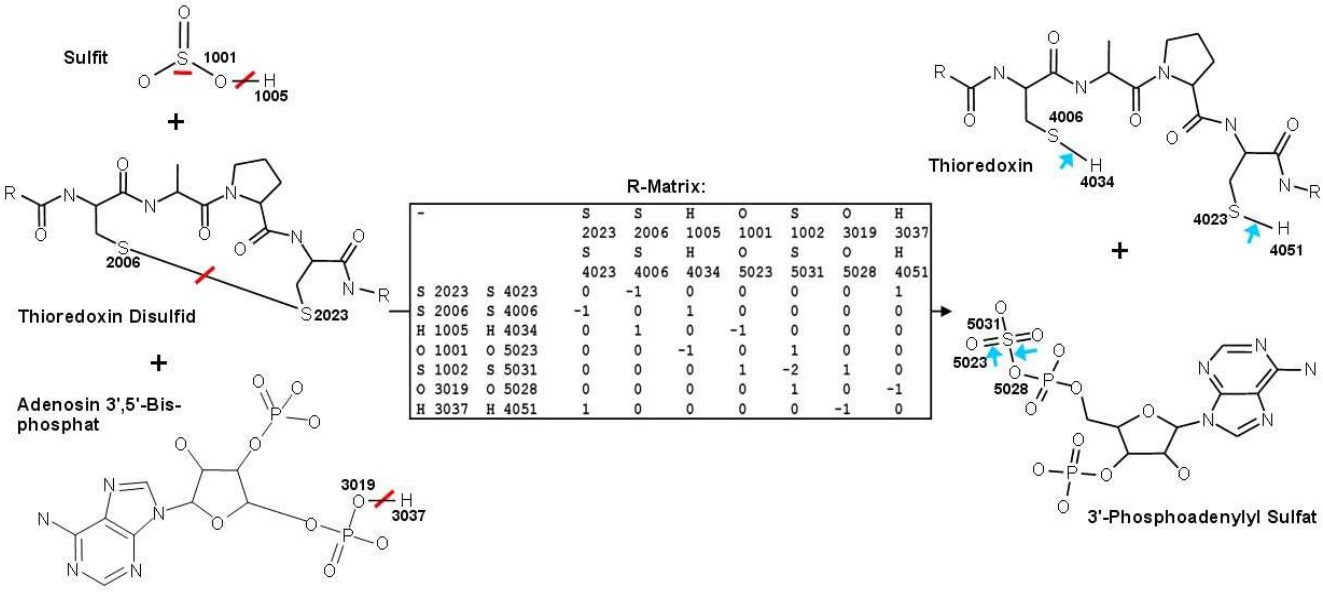


Abbildung 49: Reaktion der Glutathion-CoA-Glutathion-Transhydrogenase (EC-Nummer 1.8.4.3). Die kanonische R-Matrix dieses Enzyms ist innerhalb der Subsubklasse einmalig.



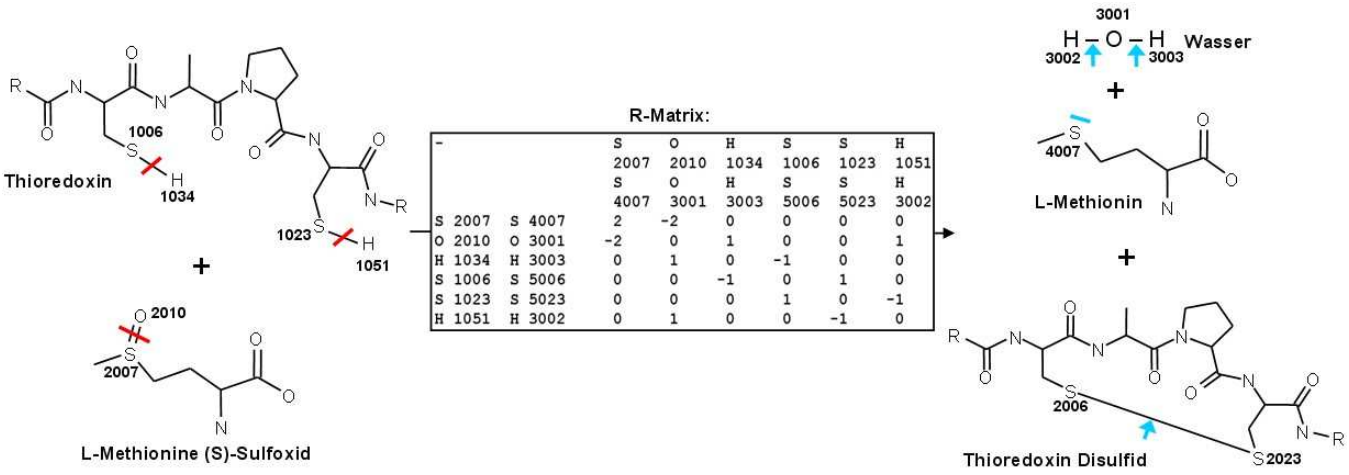
-		R 2001	H 2002	S 1021	S 1020	H 2003
		R 5001	H 3037	S 3020	S 4020	H 4037
R 2001	R 5001	2	-1	0	0	-1
H 2002	H 3037	-1	0	1	0	0
S 1021	S 3020	0	1	0	-1	0
S 1020	S 4020	0	0	-1	0	1
H 2003	H 4037	-1	0	0	1	0

Abbildung 50: R-Matrix und Reaktion der Enzym-thiol-Transhydrogenase (EC-Nummer 1.8.4.7). An der Reaktion ist ein allgemein formuliertes Akzeptormolekül beteiligt. Es wird bei Enzymen verwendet, die mehrere Elektronenakzeptoren oder Donatoren verwenden können oder bei denen die Elektronenübertragende Gruppe noch nicht definiert wurde.



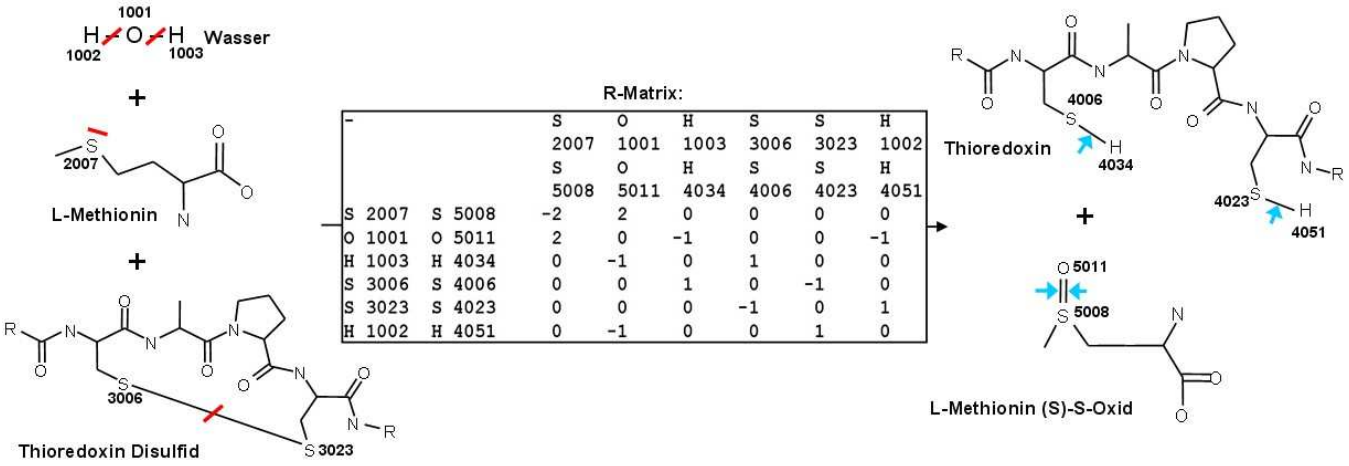
-		S 2023	S 2006	H 1005	O 1001	S 1002	O 3019	H 3037
		S 4023	S 4006	H 4034	O 5023	S 5031	O 5028	H 4051
S 2023	S 4023	0	-1	0	0	0	0	1
S 2006	S 4006	-1	0	1	0	0	0	0
H 1005	H 4034	0	1	0	-1	0	0	0
O 1001	O 5023	0	0	-1	0	1	0	0
S 1002	S 5031	0	0	0	1	-2	1	0
O 3019	O 5028	0	0	0	0	1	0	-1
H 3037	H 4051	1	0	0	0	0	-1	0

Abbildung 51: Reaktion der Phosphoadenylylsulfat-Reduktase (EC-Nummer 1.8.4.8). Die kanonische R-Matrix repräsentiert auch die EC-Nummern 1.8.4.9 und 1.8.4.10. Dieses Set bildet die größte Gruppe mit identischem R-String innerhalb der Subsubklasse.



-		S 2007	O 2010	H 1034	S 1006	S 1023	H 1051
		S 4007	O 3001	H 3003	S 5006	S 5023	H 3002
S 2007	S 4007	2	-2	0	0	0	0
O 2010	O 3001	-2	0	1	0	0	1
H 1034	H 3003	0	1	0	-1	0	0
S 1006	S 5006	0	0	-1	0	1	0
S 1023	S 5023	0	0	0	1	0	-1
H 1051	H 3002	0	1	0	0	-1	0

Abbildung 52: Reaktion der Peptid-Methionin (S)-S-oxid-Reduktase (EC-Nummer 1.8.4.11). Die kanonische R-Matrix repräsentiert auch die Reaktionen der Peptid-Methionin (R)-S-oxid-Reduktase (EC-Nummer 1.8.4.12).



-		S 2007	O 1001	H 1003	S 3006	S 3023	H 1002
		S 5008	O 5011	H 4034	S 4006	S 4023	H 4051
S 2007	S 5008	-2	2	0	0	0	0
O 1001	O 5011	2	0	-1	0	0	-1
H 1003	H 4034	0	-1	0	1	0	0
S 3006	S 4006	0	0	1	0	-1	0
S 3023	S 4023	0	0	0	-1	0	1
H 1002	H 4051	0	-1	0	0	1	0

Abbildung 53: Reaktion der L-Methionin (S)-S-oxid-Reduktase (EC-Nummer 1.8.4.13), die mit der L-Methionin (R)-S-oxid-Reduktase (EC-Nummer 1.8.4.14) die sechste Gruppe innerhalb der Subsubklasse bildet. Ihre R-Matrix entspricht der R-Matrix der Peptid-Methionin S-oxid-Reduktasen in umgekehrter Richtung.

der Subsubklasse dient eine Schwefelgruppe als Elektronendonator und Akzeptor. Bei allen Reaktionen werden Disulfidbindungen gespalten oder gebildet. Insgesamt zeigen sich aber große Abweichungen bei den an der Reaktion beteiligten Bindungen. Die Elektronen können aus SH-Gruppen, Phosphatgruppen, Sulfit- und Wassermolekülen oder anderen Donatoren stammen. Die größte Gruppe enthält 3 enzymatische Reaktionen bestehend aus den EC-Nummern 1.8.4.8, 1.8.4.9 und 1.8.4.10 (Abbildung 51). Zwei enzymatische Reaktionen (1.8.4.3 und 1.8.4.7) besitzen

eindeutige R-Strings (Abbildung 49 und 50). Die restlichen 6 Reaktionen bilden Zweiergruppen (Abbildung 48, 52 und 53).

Eine andere heterogene Gruppe unter den Oxidoreduktasen ist die Subsubklasse 1.7.1.- mit einer R-Stringidentität von 33,33 Prozent. Bei fast allen Reaktionen dieser Subsubklasse entsteht eine N-O Bindung unter Reduktion von NAD^+ oder $NADP^+$. Das Datenset dieser Subsubklasse beinhaltet 6 Reaktionen, die sich fast alle in ihren Elektronentransfermustern unterscheiden. NAD^+ und $NADP^+$ führen zu identischen Einträgen in den R-Matrizen. Bei einigen Reaktionen werden aber mehrere dieser Moleküle benötigt. Bei anderen Reaktionen entsteht keine N-O Bindung oder weitere Bindungen werden gespalten oder entstehen. Die einzige Übereinstimmung in den Elektronentransfermustern besteht zwischen den Reaktionen der Hydroxylamin-Reduktase (Abbildung 54) und N-Hydroxy-2-acetamidofluoren-Reduktase (Abbildung 55).

Wie die Subsubklassen 1.7.1.- und 1.8.4.- wurden auch die übrigen heterogenen Subsubklassen genauer analysiert. Die teils große Varianz in den Elektronentransfermustern lässt sich immer auf signifikante Unterschiede in den Reaktionen zurückführen. Eine Auflistung der R-Stringidentität der verschiedenen Subsubklassen findet sich in Tabelle 5.

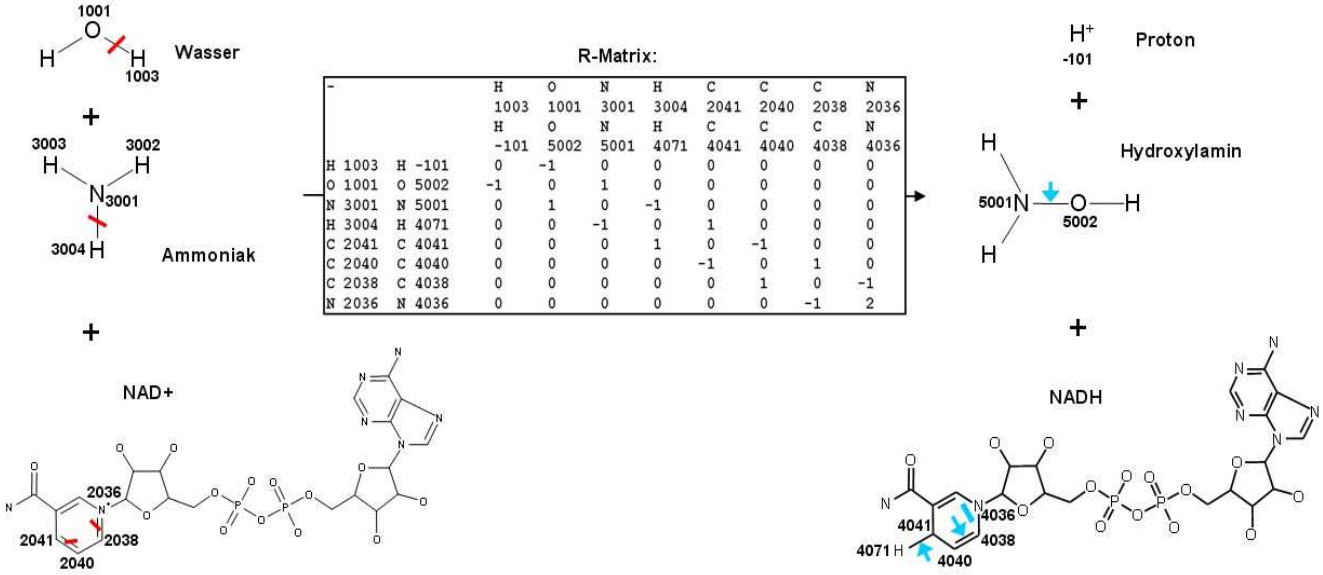


Abbildung 54: Reaktion der Hydroxylamin-Reduktase (EC-Nummer 1.7.1.10). Wie bei allen Reaktionen der Subsubklasse 1.7.1.- kommt es zur Entstehung einer Stickstoff-Sauerstoffbindung, wobei NAD+ oder NADP+ reduziert wird.

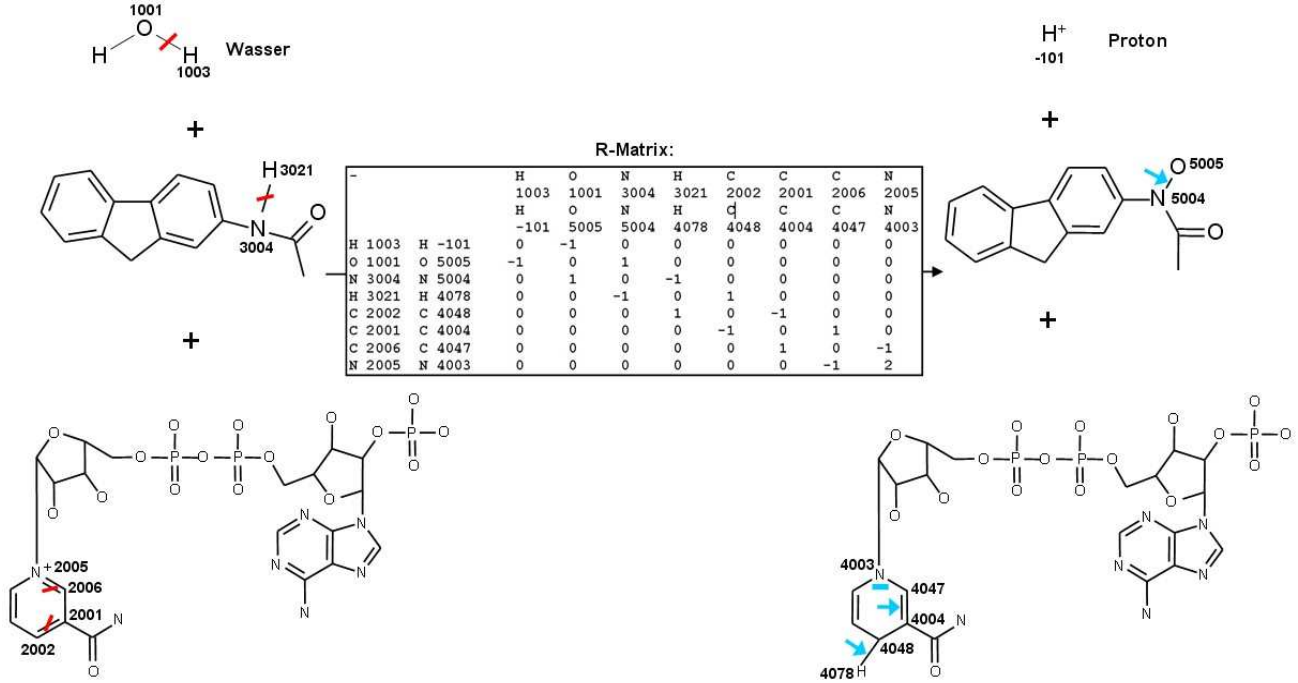


Abbildung 55: Reaktion der N-Hydroxy-2-acetamidofluoren-Reduktase (EC-Nummer 1.7.1.12). Wie bei der Reaktion der Hydroxylamin-Reduktase wird unter der Reduktion von NAD+ eine C-O und O-H Bindung gespalten. Dabei entsteht eine N-O Bindung und ein Proton wird freigesetzt.

Nur bei wenigen Reaktionen sind Fehler in den Eingangsdaten die Ursache für die Heterogenität der Subsubklasse. Hierzu gehören vor allem Reaktionen der 5. Hauptklasse, deren Subsubklassen die niedrigste Homogenität in den Elektronentransfermustern aufweisen. Bei den Isomerasereaktionen werden im Unterschied zu den übrigen Hauptklassen stereochemische Veränderungen berücksichtigt. Hierzu werden Einträge in den Molfiles berücksichtigt, welche die absolute Konfiguration der Atome kodieren. Ob ein Atom eine R-, S- oder keine Konfiguration besitzt, ist dabei aber nicht immer korrekt in den Molfiles eingetragen. Die absolute Konfiguration wird durch das Zeichenprogramm MDL ISIS Draw automatisch errechnet, wenn das Molekül gezeichnet und im Molfileformat abgespeichert wird. Entscheidend für die absolute Konfiguration ist, ob eine Bindung aus der Molekülebene hervorsteht oder hinter die Ebene zurück tritt. Hierbei können leicht Fehler entstehen, die aber kaum erkannt werden, da die absolute Konfiguration nur in sehr wenigen Anwendungen benötigt wird.

Ein Beispiel hierfür gibt die Reaktion der L-Ribulose-5-phosphat 4-Epimerase, in der L-Ribulose 5-phosphat in die epimere Verbindung D-Xylulose 5-phosphat umgewandelt wird. Beide Moleküle unterscheiden sich in der räumlichen Anordnung einer Hydroxylgruppe an einem Kohlenstoffatom,

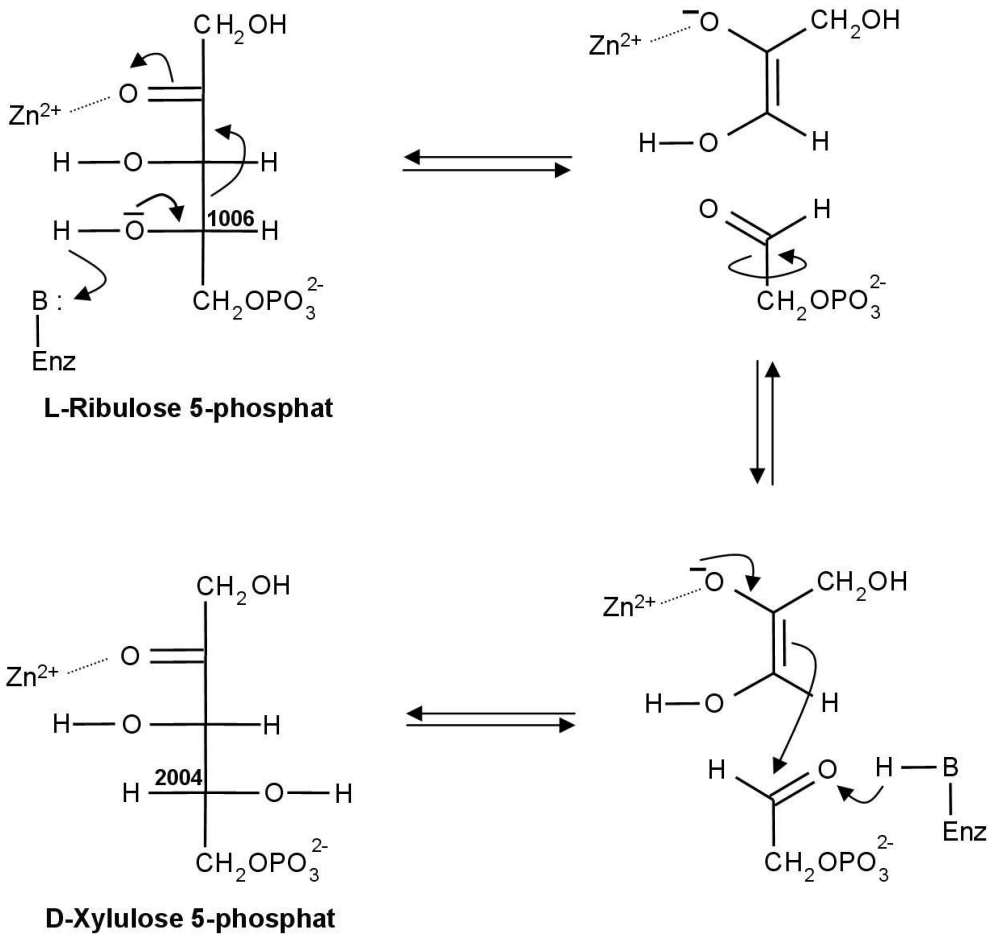


Abbildung 56: Ein Reaktionsmechanismus für die L-Ribulose-5-phosphat 4-Epimerase (EC-Nummer 5.1.3.4). Durch die Reaktion verändert sich die Orientierung der Hydroxylgruppe an dem Kohlenstoffatom C 1006 von L-Ribulose 5-phosphat, wodurch D-Xylulose 5-phosphat entsteht.

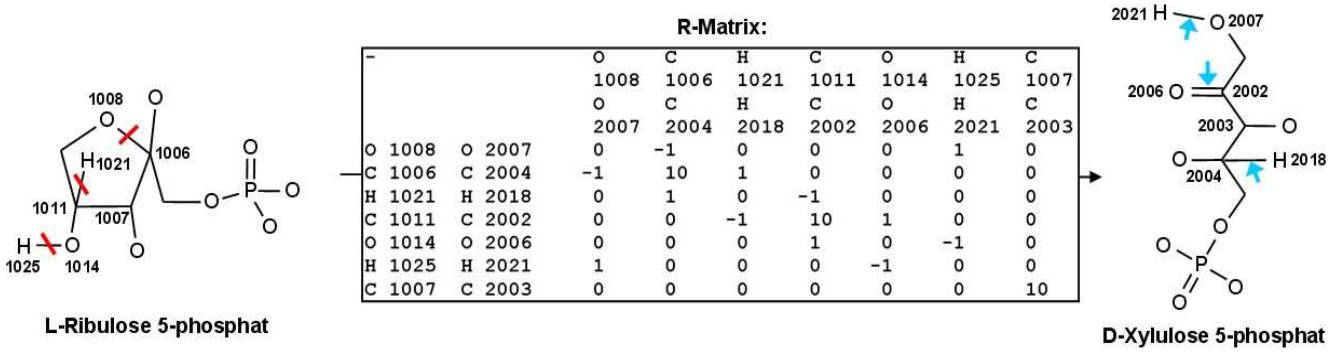


Abbildung 57: Die Eingabemoleküle L-Ribulose-5-phosphat und D-Xylulose-5-phosphat des Datensets unterscheiden sich so sehr, dass die errechnete R-Matrix für die L-Ribulose-5-phosphat 4-Epimerase zahlreiche weitere Änderungen anzeigt.

Subsub-klasse	Reaktionen/ Subsubklasse	Identität (%)	Subsub-klasse	Reaktionen/ Subsubklasse	Identität (%)	Subsub-klasse	Reaktionen/ Subsubklasse	Identität (%)
1.1.1.-	265	86	1.7.2.-	1	100	1.14.20	1	100
1.1.2.-	4	100	1.7.3.-	1	100	1.14.21	6	50
1.1.3.-	24	76	1.7.7.-	1	100	1.14.99	26	42
1.1.4.-	2	50	1.7.99.-	5	20	1.15.1	2	50
1.1.5.-	1	100	1.8.1.-	11	73	1.16.1.-	7	33
1.1.99.-	27	85	1.8.2.-	1	100	1.16.3.	1	100
1.2.1.-	63	73	1.8.3.-	4	40	1.16.8	1	100
1.2.2.-	3	66	1.8.4.	11	27	1.17.1	4	75
1.2.3.-	6	50	1.8.5.	1	100	1.17.3	2	50
1.2.4.-	3	100	1.8.7	1	100	1.17.4	2	100
1.2.7.-	8	62	1.8.98	1	100	1.17.5	1	100
1.2.99.-	6	50	1.8.99	1	100	1.17.99	4	50
1.3.1.-	67	59	1.9.3.	1	100	1.18.1.	4	50
1.3.2.-	1	100	1.9.6.	1	100	1.18.6	1	50
1.3.3.-	9	33	1.9.99	1	100	1.19.6.	1	100
1.3.5.-	1	100	1.10.1	1	100	1.20.1.-	1	100
1.3.7.-	2	50	1.10.2	2	50	1.20.4	2	100
1.3.99.-	17	47	1.10.3	3	33	1.20.98	1	100
1.4.1.-	20	80	1.10.99	3	33	1.20.99	1	100
1.4.2.-	1	100	1.11.1	14	18	1.21.3	6	33
1.4.3.-	15	93	1.12.1	2	100	1.21.4	3	67
1.4.4.-	1	100	1.12.2	1	100	1.21.99	1	100
1.4.7.-	1	100	1.12.5	1	100	1.97.1	6	25
1.4.99.-	5	67	1.12.7	1	100	2.1.1	106	42
1.5.1.-	26	38	1.12.98.	2	50	2.1.2	2	100
1.5.3.-	9	56	1.12.99.	1	100	2.1.3.	8	75
1.5.4.-	1	100	1.13.11	47	26	2.1.4	1	100
1.5.5.-	1	100	1.13.12	5	40	2.2.1	8	50
1.5.7.-	1	100	1.13.99	2	50	2.3.1	131	38
1.5.8.-	2	100	1.14.11	22	59	2.3.2	4	50
1.5.99.-	10	40	1.14.12	16	56	2.3.3	13	86
1.6.1.-	2	100	1.14.13	91	63	2.4.1	129	73
1.6.2.-	1	100	1.14.14	2	50	2.4.2	21	46
1.6.3.-	1	100	1.14.15.	7	43	2.4.99	7	86
1.6.5.-	4	100	1.14.16	5	40	2.5.1	43	14
1.6.6.-	1	100	1.14.17	3	33	2.6.1	74	55
1.6.99.-	4	100	1.14.18	2	50	2.6.3	1	100
1.7.1.-	6	33	1.14.19	2	100	2.6.99	1	50

Tabelle 5a: R-Stringidentität der Subsubklassen.

Subsub-klasse	Reaktionen/ Subsubklasse	Identität (%)	Subsub-klasse	Reaktionen/ Subsubklasse	Identität (%)	Subsub-klasse	Reaktionen/ Subsubklasse	Identität (%)
2.7.1	109	82	3.2.1	31	63	4.1.99	4	50
2.7.2	13	92	3.2.2	12	67	4.2.1	86	42
2.7.3	8	88	3.3.1	2	50	4.2.2	1	100
2.7.4	19	84	3.3.2	9	56	4.2.3	21	17
2.7.6	5	80	3.4.11	20	100	4.2.99	1	100
2.7.7	42	54	3.4.13	12	100	4.3.1	16	38
2.7.8	11	78	3.4.14	9	100	4.3.2	2	50
2.7.9	5	25	3.4.15	3	100	4.3.3	2	50
2.7.10	2	100	3.4.16	4	100	4.4.1	15	19
2.7.11	28	100	3.4.17	19	100	4.5.1	4	25
2.7.12	2	100	3.4.18	1	100	4.6.1	6	67
2.7.13	2	100	3.4.19	9	100	4.99.1	3	100
2.7.99	1	100	3.4.21	98	100	5.1.1	15	47
2.8.1	5	40	3.4.22	50	100	5.1.2	6	100
2.8.2	23	91	3.4.23	38	100	5.1.3	14	18
2.8.3	14	50	3.4.24	80	100	5.1.99	4	67
2.8.4	1	100	3.4.25	1	100	5.2.1	10	100
2.9.1	1	100	3.5.1	76	81	5.3.1	21	52
3.1.1	57	82	3.5.2	15	71	5.3.2	2	100
3.1.2	20	100	3.5.3	19	68	5.3.3	14	38
3.1.3	61	71	3.5.4	24	58	5.3.99	7	14
3.1.4	6	100	3.5.5	6	80	5.4.1	1	100
3.1.5	1	100	3.5.99	5	40	5.4.2	10	20
3.1.6	10	90	3.6.1	30	83	5.4.3	7	29
3.1.7	2	100	3.6.2	2	100	5.4.4	3	67
3.1.8	1	100	3.6.3	44	97	5.4.99	14	7
3.1.11	6	100	3.6.4	11	100	5.5.1	12	50
3.1.13	5	100	3.6.5	6	100	5.99.1	1	100
3.1.14	1	100	3.7.1	10	80	6.1.1	21	90
3.1.15	1	100	3.8.1	6	50	6.2.1	29	100
3.1.16	1	100	3.9.1	1	100	6.3.1	8	87
3.1.21	7	100	3.10.1	2	100	6.3.2	17	94
3.1.22	4	100	3.11.1	2	100	6.3.3	2	50
3.1.25	1	100	3.12.1	1	100	6.3.4	8	40
3.1.26	11	100	3.13.1	1	100	6.3.5	6	33
3.1.27	10	100	4.1.1	69	80	6.4.1	5	100
3.1.30	2	100	4.1.2	30	60	6.5.1.	2	100
3.1.31	1	100	4.1.3	20	80	6.6.1	1	100

Tabelle 5b: R-Stringidentität der Subsubklassen.

wie dies auch aus dem Reaktionsmechanismus der L-Ribulose-5-phosphat 4-Epimerase (Abbildung 56) zu ersehen ist. Die Strukturen der beiden Eingabemoleküle gehen aus Abbildung 57 hervor. L-Ribulose 5-phosphat ist in der zyklischen Form, D-Xylulose 5-phosphat in der offenkettigen Form gezeichnet worden. Kohlenhydrate können durch eine intramolekulare Reaktion von der offenkettigen Form in die zyklische Form übergehen. Diese Reaktion dominiert die R-Matrix. 3 Bindungen werden gespalten und 3 Bindungen entstehen. Ferner entsteht ein neues Chiralitätszentrum. Hierdurch wird die eigentliche Epimerasereaktion überlagert.

3.4 Gruppierung der Subsubklassen nach R-String Identität

Wie in Abschnitt 3.3 beschrieben, sind die Elektronentransfermuster innerhalb der Subsubklassen sehr homogen. Eine Ausnahme bilden die enzymatischen Reaktionen der unspezifischen Subsubklassen (a.b.98 oder a.b.99) oder der 5. Hauptklasse, wo zusätzlich die absolute Konfiguration berücksichtigt wird. Dennoch beträgt die durchschnittliche R-Stringidentität innerhalb der Subsubklassen immerhin 76 Prozent. Um die Reaktionen der verschiedenen Subsubklassen besser vergleichen zu können, wurde daher für jede Subsubklasse der häufigste R-String errechnet und als repräsentativ für die jeweilige Subsubklasse betrachtet.

Die R-Strings der verschiedenen Subsubklassen wurden verglichen und Subsubklassen mit identischen Elektronentransfermustern gruppiert. Von den 228 untersuchten Subsubklassen besitzen 121 ein spezifisches Elektronentransfermuster. Die R-Strings dieser Subsubklassen lassen sich nicht gruppieren. Die übrigen 107 Subsubklassen bilden hingegen 26 Gruppen unterschiedlicher Größe. Tabelle 6 zeigt die verschiedenen Gruppen von Subsubklassen mit identischen R-Strings. Neben jeder Gruppe sind die Veränderungen aufgelistet, die das Elektronentransfermuster kennzeichnen. Hierzu gehören die Bindungen, die auf der Eduktseite gespalten werden oder auf der Produktseite entstehen. Nimmt die Anzahl der Valenzelektronen eines Atoms ab oder zu, so wird dies durch Punkte neben dem Atomsymbol gekennzeichnet. Dabei repräsentiert jeder Punkt ein Elektron.

Die Gruppen wurden nach ihrer Größe sortiert und nummeriert. Die größten Gruppen setzen sich aus Subsubklassen zusammen, deren R-Matrizen die Dimension 4 besitzen. Wie aus Tabelle 4 hervorgeht, ist diese Matrixdimension unter den Subsubklassen am häufigsten zu finden. Bei diesem Reaktionstyp werden fast immer 2 Bindungen auf der Eduktseite gespalten, während auf der Produktseite 2 Bindungen neu entstehen. Die Subsubklassen der größeren Gruppen stammen

Gruppen Nummer	Subsub-klassen	gespaltene Bindungen / abgegebene Elektronen	neue Bindungen / aufgenommene Elektronen
1	2.4.1.- 2.4.99.- 3.1.1.- 3.1.3.- 3.1.4.- 3.1.5.- 3.1.7.- 3.1.8.- 3.1.11.- 3.1.13.- 3.1.15.- 3.1.16.- 3.1.21.- 3.1.25.- 3.1.26.- 3.1.30.- 3.2.1.- 3.3.2.-	C-O, O-H	C-O, O-H
2	2.4.2.- 3.2.2.- 3.4.11.- 3.4.13.- 3.4.14.- 3.4.15.- 3.4.16.- 3.4.17.- 3.4.18.- 3.4.19.- 3.4.21.- 3.4.22.- 3.4.23.- 3.4.24.- 3.4.25.- 3.5.1.- 3.5.2.-	C-N, O-H	C-O, N-H
3	2.7.1.- 2.7.2.- 2.7.4.- 2.7.6.- 2.7.8.- 2.7.10.- 2.7.11.- 2.7.12.- 2.7.99.- 3.6.1.- 3.6.3.- 3.6.4.- 3.6.5.- 4.6.1.-	P-O, O-H	P-O, O-H
4	3.1.14.- 3.1.22.- 3.1.27.- 3.1.31.-	P-O, C-O, O-H, O-H	P-O, C-O, O-H, O-H
5	3.7.1.- 4.1.1.- 4.1.2.- 4.1.3.-	C-C, O-H	C-O, C-H
6	5.1.1.- 5.1.2.- 5.1.3.- 5.1.99.-	S-Konfiguration	R-Konfiguration
7	1.1.1.- 1.2.1.- 1.17.1.-	C-N, C-C, C-H, C-H	C-O, C-C, C-H / N:
8	1.1.99.- 1.2.99.- 1.17.99.-	C-H, O-H	C-O, R-H
9	2.3.1.- 3.1.2.- 3.3.1.-	S-C, O-H	C-O, S-H
10	2.8.2.- 3.1.6.- 3.6.2.-	S-O, O-H	S-O, O-H
11	4.3.1.- 4.3.2.- 4.3.3.-	C-N, C-H	C-C, N-H
12	6.3.1.- 6.3.2.- 6.3.3.-	P-O, C-O, N-H	P-O, N-C, O-H
13	1.1.2.- 1.2.2.-	O-H, C-H	C-O / Fe•
14	1.1.3.- 1.2.3.-	O-O, O-H, C-H	C-O, O-H, O-H
15	1.2.4.- 1.4.4.-	S-S, C-C, O-H	S-C, O-C, S-H
16	1.4.1.- 1.5.1.-	C-C, C-N, C-N, O-H, O-H, C-H	C=O, C-C, N-H, C-H / N:
17	1.4.99.- 1.5.99.-	C-N, O-H, O-H, C-H	C=O, N-H, R-H
18	1.5.3.- 1.17.3.-	C-N, C-H, O-O, O-H, O-H	C=O, N-H, O-H, O-H
19	1.16.1.- 1.18.1.-	N-C, C-C, Fe• , Fe•	C-C, C-H / N:
20	1.17.4.- 1.17.5.-	S-S, O-H, C-H	C-O, S-H, S-H
21	2.1.4.- 2.3.2.-	C-N, N-H	C-N, N-H
22	2.7.3.- 2.7.13.-	P-O, N-H	P-N, O-H
23	3.5.3.- 3.5.4.-	N-C, N-C, O-H, O-H	C=O, N-H, N-H
24	4.2.1.- 4.2.99.-	O-C, C-H	C-C, O-H
25	5.2.1.- 5.3.3.-	C-C, C-H	C-C, C-H
26	5.3.2.- 5.5.1.-	C-O, C-H	C-C, O-H

Tabelle 6: Vergleich und Gruppierung der Subsubklassen nach ihren R-Strings führt zu 26 Gruppen, die jeweils einen identischen R-String aufweisen. Die 26 Gruppen setzen sich aus 107 Subsubklassen zusammen. Die Enzyme innerhalb einer Gruppe können unterschiedlichen Hauptklassen des EC-Klassifikationssystems angehören. Dennoch besitzen sie identische Elektronentransfermuster.

überwiegend aus den Hauptklassen der Transferasen, Hydrolasen und Lyasen. Die größte Gruppe setzt sich aus 18 Subsubklassen zusammen. Es handelt sich überwiegend um Subsubklassen aus der Hauptklasse der Hydrolasen. Daneben sind 2 Subsubklassen aus der zweiten Hauptklasse beteiligt. Die Reaktion der Glykogen Synthase aus Abbildung 58, bei der eine Glukoseeinheit übertragen wird, gehört zu dieser Gruppe. Bei allen Reaktionen der größten Gruppe wird auf der Eduktseite eine C-O und O-H Bindung gespalten, während auf der Produktseite wiederum eine C-O und O-H Bindung entsteht. Diese Bindungstypen treten häufig in Molekülen auf und sind oft an Reaktionen beteilig. Hierzu tragen vor allem die Sauerstoffatome bei, einerseits mit ihren freien Elektronenpaaren, andererseits mit ihrer Eigenschaft, Bindungselektronen an sich zu ziehen. Das Elektronentransfermuster ist vergleichsweise einfach. Es tritt bei vielen Spaltungsreaktionen auf, insbesondere wenn Wassermoleküle an der Spaltung beteiligt sind. Daher ist die Anhäufung der Hydrolasereaktionen in Gruppe 1 nicht verwunderlich. Bei den beiden Subsubklassen 2.4.1.- und 2.4.99.- aus der Hauptklasse der Transferasen handelt es sich um Übertragungsreaktionen, wobei ein Teil eines Moleküls auf ein anderes Molekül übertragen wird.

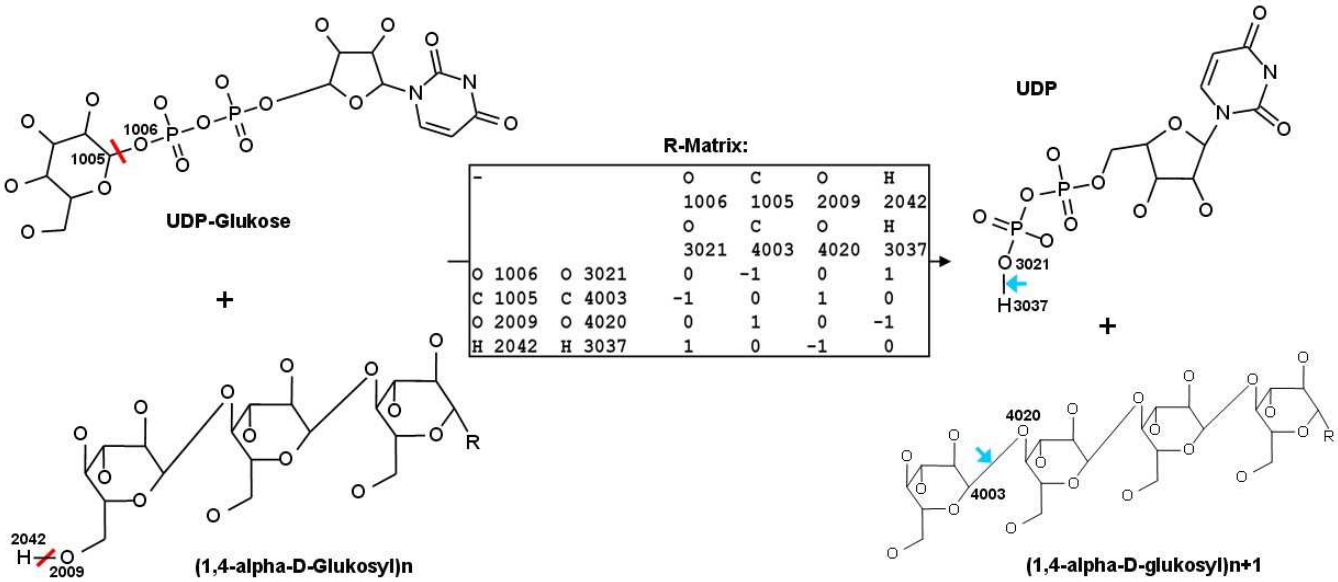


Abbildung 58: Reaktion und R-Matrix der Glykogen (Stärke) Synthase (EC-Nummer 2.4.1.11). Eine Glucoseeinheit wird von UDP-Glukose abgespalten und auf Glykogen übertragen. Glykogen ist ein Polysaccharid aus Glucose-Einheiten. Die Reaktion repräsentiert nicht nur die Subsubklasse 2.4.1.-, sondern auch weitere 17 Subsubklassen mit identischem Elektronentransfermuster.

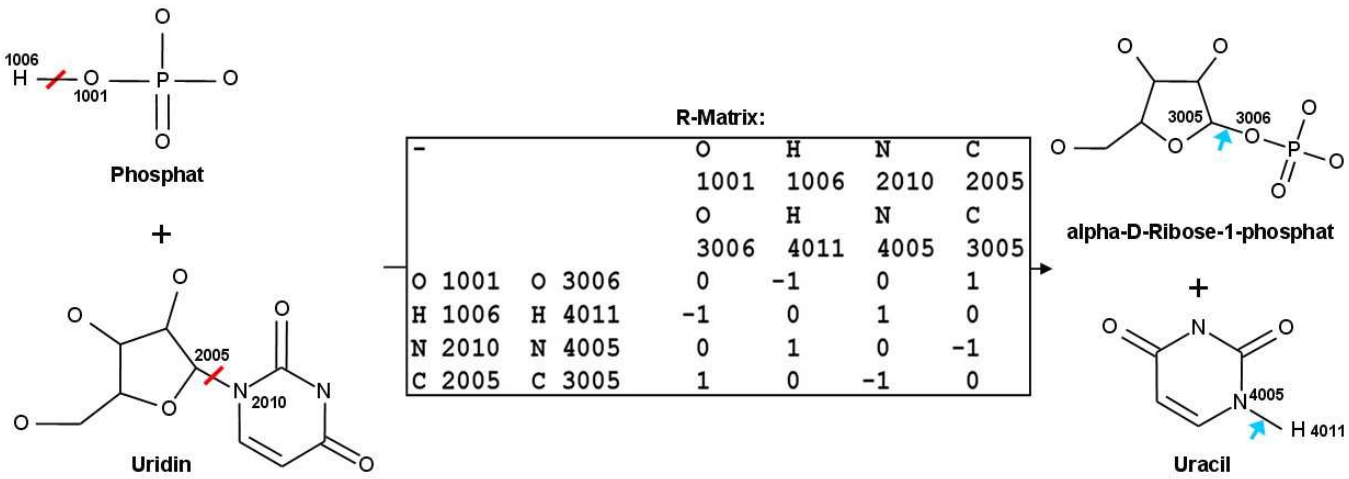


Abbildung 59: Reaktion und R-Matrix der Uridin-Phosphorylase (EC-Nummer 2.4.2.3). Die Reaktion repräsentiert die zweitgrößte Gruppe von Subsubklassen mit identischem Elektronentransfermuster.

Bei der Reaktion der Uridin-Phosphorylase (Abbildung 59) wird auf der Eduktseite eine C-N und O-H Bindung gespalten. Auf der Produktseite entsteht andererseits eine C-O und N-H Bindung. Dieses Muster tritt bei allen 17 Subsubklassen der zweitgrößten Gruppe auf. Die zweitgrößte Gruppe setzt sich ebenfalls aus Transferasen und Hydrolasen zusammen.

Am heterogensten in Bezug auf die Hauptklassenzusammensetzung ist Gruppe 3, die enzymatische Reaktionen aus der zweiten, dritten und vierten Hauptklasse enthält. Die Abbildungen 60 bis 62 zeigen je eine Reaktion aus einer der drei verschiedenen Hauptklassen. Bei allen Reaktionen wird auf der Eduktseite eine P-O und O-H Bindung gespalten, während auf der Produktseite eine P-O und O-H Bindung entsteht. Bei allen Reaktionen sind Phosphatgruppen beteiligt, die aufgespalten werden. Die chemischen Eigenschaften der Molekülstrukturen, die am Reaktionskern beteiligt sind, sind weitgehend identisch.

Unter den Gruppen von Subsubklassen mit identischen Elektronentransfermustern gibt es insgesamt 13 Gruppen, die eine R-Matrixdimension von 4 besitzen und wo auf der Eduktseite 2 Bindungen gespalten werden, während auf der Produktseite 2 Bindungen entstehen. Dieses Muster ist typisch für Reaktionen, bei denen 2 Moleküle Atome oder funktionelle Gruppen untereinander austauschen.

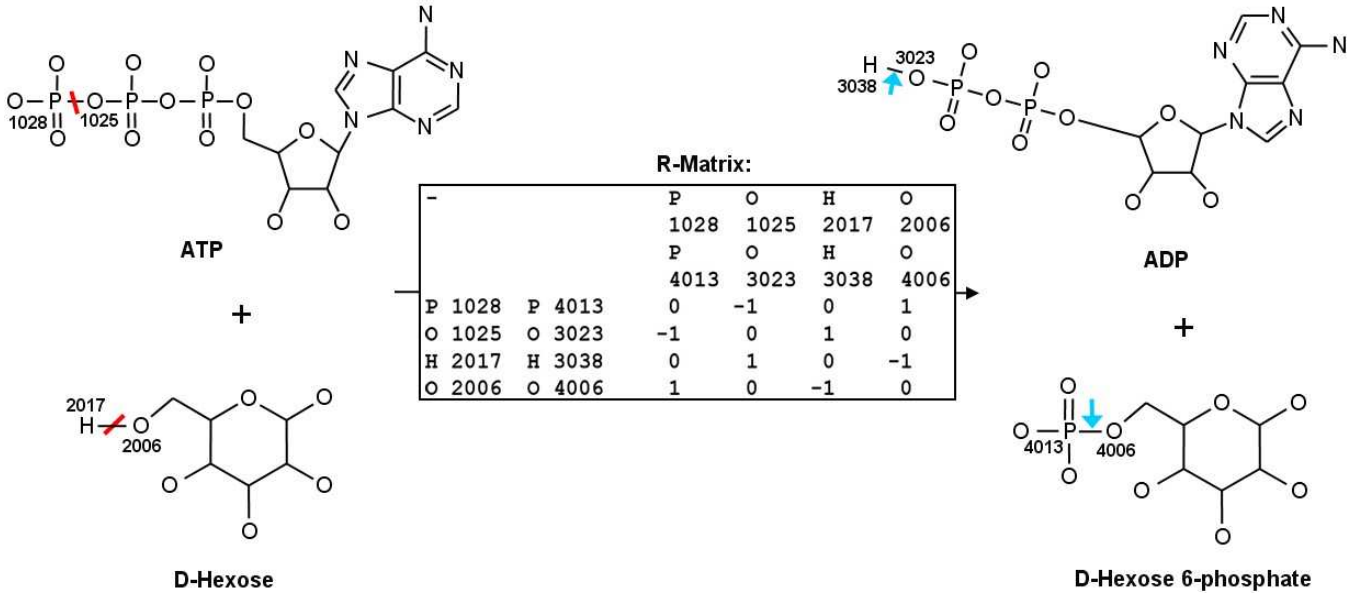


Abbildung 60: Reaktion und R-Matrix der Hexokinase (EC-Nummer 2.7.1.1). Unter Verbrauch von ATP wird D-Hexose zu D-Hexose 6-phosphat phosphoryliert.

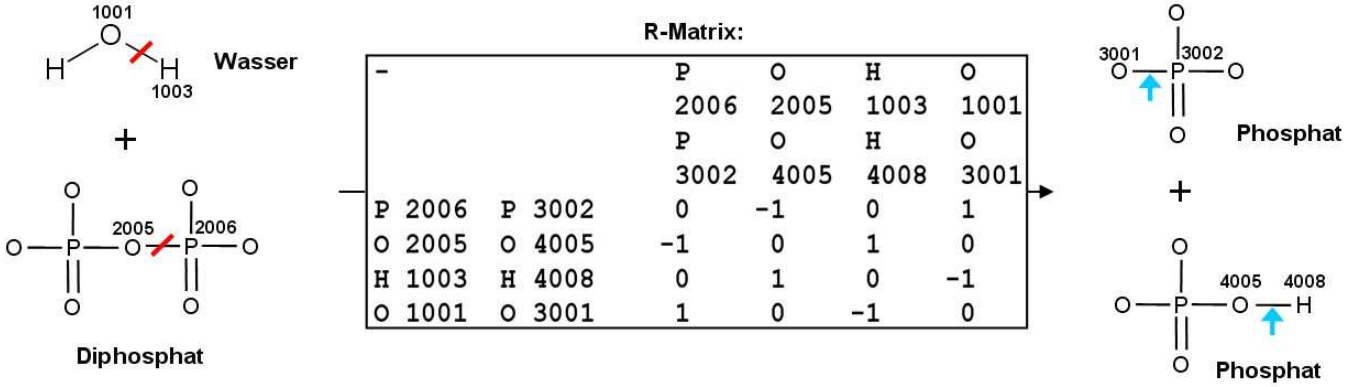


Abbildung 61: Reaktion und R-Matrix der Inorganischen Diphosphatase (EC-Nummer 3.6.1.1), die Diphosphat unter der Beteiligung von Wasser in 2 Phosphatmoleküle spaltet.

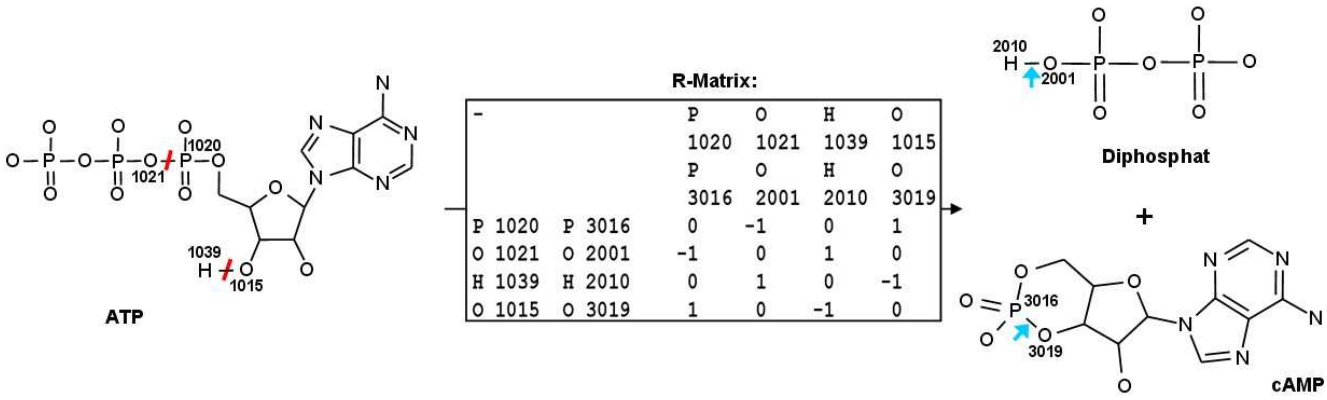


Abbildung 62: R-Matrix und Reaktion der Adenylat-Zyklase (EC-Nummer 4.6.1.1), die unter Abspaltung von Diphosphat das zyklische Adenosinmonophosphat (cAMP) erzeugt. Die Reaktion repräsentiert die 6 Reaktionen der Subsubklasse 4.6.1.-. Der Reaktionskern besitzt das gleiche Elektronentransfermuster wie die Reaktionen der Subsubklassen 2.7.1.- und 3.6.1.-.

Eine Ausnahme unter den Gruppen bilden die Isomerase-Reaktionen der Gruppe 6. Bei dieser Gruppe ändert sich lediglich die absolute Konfiguration eines Atoms von S nach R. Die zugehörige R-Matrix besitzt die Dimension 1 und beinhaltet einen Matrixeintrag von 20. Zu der Gruppe gehört ebenfalls die Alanin-Racemase (EC-Nummer 5.1.1.1), deren Reaktion und R-Matrix in den Abbildungen 35 und 36 (Abschnitt 2.3.1) dargestellt werden.

Die Isomerasereaktionen der Gruppen 25 und 26 beinhalten hingegen keine Verränderungen der absoluten Konfiguration. Es handelt sich hierbei um intramolekulare Veränderungen, die Bindungsumbrüche beinhalten. Hierzu gehören intramolekulare Oxidoreduktasen und Transferasen. Die R-Matrix der Phenylpyruvat-Tautomerase (Abbildung 63) repräsentiert die Reaktionen der Gruppe 26.

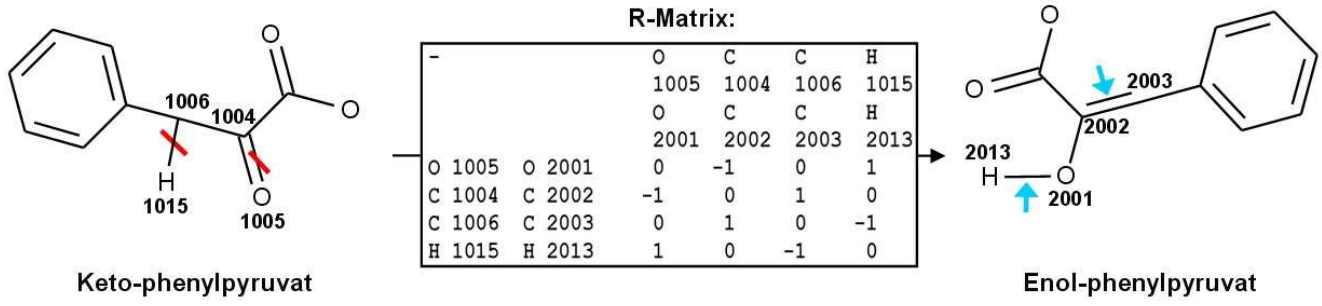
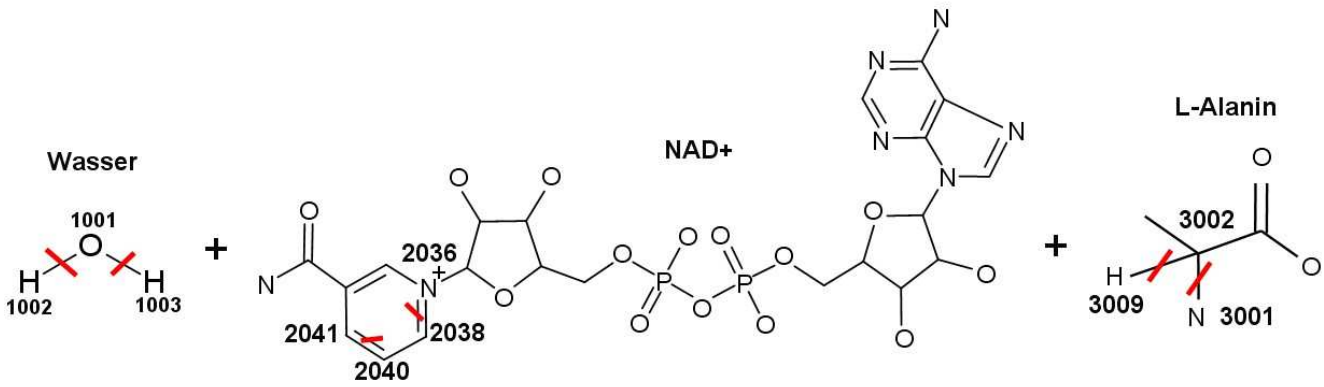


Abbildung 63: Reaktion und R-Matrix der Phenylpyruvat-Tautomerase (EC-Nummer 5.3.2.1). Die Umwandlung von Keto-Phenylpyruvat zu Enol-phenylpyruvat führt zu keiner Veränderung der absoluten Konfiguration.



R-Matrix:

-		H 1003	O 1001	H 1002	N 3001	C 3002	H 3009	C 2041	C 2040	C 2038	N 2036
		H -101	O 6003	H 5004	N 5001	C 6002	H 4071	C 4041	C 4040	C 4038	N 4036
H 1003	H -101	0	-1	0	0	0	0	0	0	0	0
O 1001	O 6003	-1	0	-1	0	2	0	0	0	0	0
H 1002	H 5004	0	-1	0	1	0	0	0	0	0	0
N 3001	N 5001	0	0	1	0	-1	0	0	0	0	0
C 3002	C 6002	0	2	0	-1	0	-1	0	0	0	0
H 3009	H 4071	0	0	0	0	-1	0	1	0	0	0
C 2041	C 4041	0	0	0	0	0	1	0	-1	0	0
C 2040	C 4040	0	0	0	0	0	0	-1	0	1	0
C 2038	C 4038	0	0	0	0	0	0	0	1	0	-1
N 2036	N 4036	0	0	0	0	0	0	0	0	-1	2

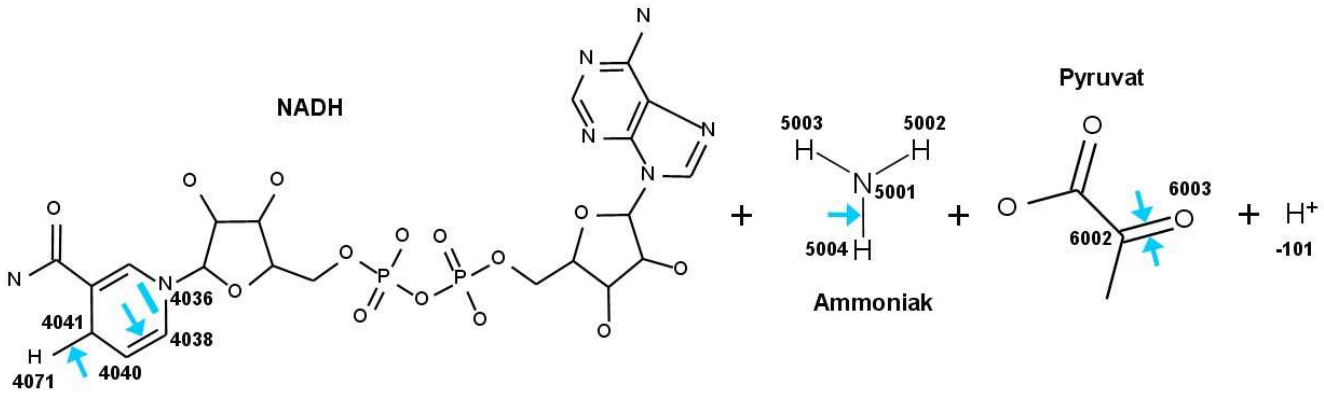


Abbildung 64: Die Alanin-Dehydrogenase oxidiert L-Alanin unter der Beteiligung eines Wassermoleküls und NAD+ zu Pyruvat. Sie repräsentiert die Subsubklasse 1.4.1.-, die mit der Subsubklasse 1.5.1.- die Gruppe mit der komplexesten R-Matrix bildet.

Die Gruppe 16 wird aus den Subsubklassen 1.4.1.- und 1.5.1.- gebildet. Die R-Matrix der Gruppe beinhaltet die zahlreichsten Veränderungen unter den Subsubklassen, die an der Gruppenbildung beteiligt sind. An allen Reaktionen der Gruppe 16 sind die Kofaktoren NAD oder NADP beteiligt. Hierdurch entstehen über die eigentlichen Zielmoleküle der Reaktionen hinaus die gleichen Elektronentransfermuster. Da bei vielen Reaktionen der Oxidoreduktasen Kofaktoren beteiligt sind, besitzen sie häufig größere Matrixdimensionen als die R-Matrizen anderer Hauptklassen. Daher sind Oxidoreduktase Reaktionen nicht an den größten Gruppen beteiligt und lassen sich auch mit keiner Subsubklasse anderer Hauptklassen gruppieren.

Die Oxidoreduktasen stellen hingegen den größten Anteil unter den Subsubklassen, die ein spezifisches Elektronentransfermuster besitzen und sich nicht anhand ihrer R-Strings gruppieren lassen.

4. Diskussion

Das im Rahmen dieser Arbeit entwickelte Verfahren ermöglicht eine automatisierte Charakterisierung von enzymatischen Reaktionen auf Grundlage des Dugundji-Ugi-Modells. Das Elektronen-austauschmuster einer Reaktion wird hierbei durch Reaktionsmatrizen (R-Matrizen) ausgedrückt, die Aufschluss über die Entstehung oder Spaltung von chemischen Bindungen geben.
Die manuelle Konstruktion von R-Matrizen ist sehr zeitaufwendig und erlaubt nur die Betrachtung eines begrenzten Sets von enzymatischen Reaktionen (Svena, 2002). Das Hauptziel dieser Arbeit bestand daher zunächst darin, die Vorraussetzung dafür zu schaffen, dass R-Matrizen in großer Anzahl automatisch berechnet werden können. In einem zweiten Schritt wurde das Programm zur R-Matrix-Berechnung auf ein größeres Datenset von biochemischen Reaktionen angewendet. Von besonderem Interesse waren hierbei die Charakterisierung und der Vergleich der Reaktionen anhand ihres Elektronentransfermusters.

4.1 Automatisierte R-Matrix-Berechnung

R-Matrizen lassen sich vergleichsweise einfach errechnen, wenn eine Zuordnung der Eduktatome auf die Produktatome vorliegt. Ohne Atomzuordnung besteht hingegen kein nachvollziehbarer Reaktionsablauf. Die Lösung des Atomzuordnungsproblems erwies sich als sehr komplex und vereinnahmte einen großen Teil des Projektes. Der neu entwickelte Algorithmus zur Bestimmung der größten gemeinsamen Substruktur ist der Kern des Verfahrens. Daneben war die Implementierung weiterer Module erforderlich, beispielsweise zur Bewertung der größten gemeinsamen Teilstrukturen, zur Erkennung aromatischer Ringe oder zur Errechnung und Kanonisierung der R-Matrizen.

4.1.1 MCS Algorithmen und Molekülstrukturvergleich

Das wichtigste Kernstück des neuen Verfahrens zur automatisierten R-Matrixberechnung ist der *c*-MCS-Algorithmus (vgl. 2.1). Der Algorithmus basiert auf graphentheoretischen Methoden, die zwei Graphen G_1 und G_2 vergleichen und die maximalen gemeinsamen Subgraphen (englische Bezeichnung „Maximal Common Subgraph“) der Eingabegraphen errechnen.
Da Moleküle ebenfalls Graphen repräsentieren, werden MCS Algorithmen seit vielen Jahren zum Vergleich von Molekülen verwendet (Hattori *et al.*, 2003, Marialke *et al*., 2007, Raymond *et al.*, 2002, García *et al*., 2004, Durand *et al*., 1999, McGregor *et al*., 1982, Bayada *et al*., 1992, usw.). MCS Algorithmen erlauben einen sehr präzisen Vergleich der Molekülstrukturen. Der wesentliche

Nachteil des Verfahrens besteht allerdings darin, dass die MCS-Suche zur Komplexitätsklasse der NP-vollständigen Probleme gehört. Zu diesen Problemen konnten bisher noch keine Algorithmen entwickelt werden, die das Problem in polynomieller Zeit lösen. Mit der Größe der Eingabegraphen nimmt die Laufzeit exponentiell zu. Aus diesem Grund lassen sich nicht Eingabegraphen beliebiger Größe miteinander vergleichen. Mit zunehmender Molekülgröße und in Abhängigkeit von der Molekülstruktur wird sehr schnell eine Grenze erreicht, ab der die Laufzeit unvorstellbare Dimensionen annimmt. Um dennoch Eingabegraphen ausreichender Größe miteinander vergleichen zu können, werden häufig heuristische Informationen in die MCS-Suche integriert. So kann die Betrachtung der atomaren Umgebung oder die Einbeziehung dreidimensionaler Informationen das Laufzeitverhalten erheblich verbessern. Allerdings führen diese zusätzlichen Kriterien zu weiteren Einschränkungen. Besonders bei der Betrachtung von Reaktionen können diese Kriterien zu falschen Schlussfolgerungen führen, denn die atomare Umgebung eines Atoms kann sich in einer Reaktion verändern und die dreidimensionale Struktur kann sich umformen.
Die Rekonstruktion von Reaktionen erfordert daher einen flexiblen MCS-Algorithmus, der mit wenigen Informationen auskommt, damit auch komplexe Veränderungen der Molekülstruktur nachvollziehbar sind. Ein hohes Maß an Flexibilität führt andererseits zu einer Verschlechterung des Laufzeitverhaltens. Viele enzymatische Reaktionen beinhalten aber die Umwandlung großer Metabolite. Ein entsprechender MCS-Algorithmus, der diese gegensätzlichen Eigenschaften befriedigend vereinbart, war leider nicht verfügbar. Im Rahmen dieses Projektes wurde deshalb ein neuer MCS-Algorithmus entwickelt.

4.1.2 *c*-MCS-Algorithmus

Im Rahmen dieser Arbeit wurden verschiedene MCS Algorithmen implementiert und getestet. Keiner der Algorithmen erwies sich als ausreichend, um die gestellten Anforderungen zu erfüllen. Die Algorithmen waren entweder zu langsam oder zu unflexibel. Aus dieser Situation resultierte die Idee, zwei Algorithmen mit ihren unterschiedlichen Eigenschaften miteinander zu verknüpfen. Der erste schnellere Algorithmus basiert auf einer Variante des Bron-Kerbosch-Algorithmus (Bron und Kerbosch, 1973). Bei dem zweiten Algorithmus handelt es sich um den so genannten McGregor-Algorithmus (McGregor, 1982), der sehr flexibel ist und für den Vergleich von Molekülen entwickelt wurde.

4.1.2.1 Variante des Bron-Kerbosch-Algorithmus

Der Bron-Kerbosch-Algorithmus und seine Varianten lösen das MCS-Problem durch Überführung in das Cliquen-Problem (vgl. 2.1.3). Ausgangspunkt sind die beiden Eingabegraphen, die in diesem Fall durch zwei Moleküle repräsentiert werden. Aus den beiden Eingabemolekülen wird zunächst ein übergeordneter Produktgraph gebildet, der die kompatiblen Bereiche der Moleküle widerspiegelt. Sind im Produktgraph zwei Knoten über eine Kante miteinander verbunden, so ist dies ein Indiz für eine strukturelle Gemeinsamkeit der beiden Moleküle. Vollständige Teilgraphen innerhalb des Produktgraphen, die so genannten Cliquen, korrespondieren mit gemeinsamen Subgraphen (bzw. Substrukturen) der Moleküle. Maximale Cliquen korrespondieren daher mit maximalen gemeinsamen Subgraphen (Levi, 1972).

Die Suche nach den maximalen Cliquen innerhalb des Produktgraphen erfolgt mit Hilfe von Rücksetzverfahren, die auch als Backtracking (engl., Rückverfolgung) Verfahren bezeichnet werden. Der ursprüngliche Bron-Kerbosch-Algorithmus unterscheidet nicht zwischen den beiden Kantentypen des Produktgraphen. Die schnellere Variante hingegen unterscheidet zwischen den *c*- und *d*-Kanten. Die *d*-Kanten zeigen an, dass die Atome der beiden Eingabemoleküle keine strukturelle Ähnlichkeit zueinander aufweisen. Die *c*-Kanten („*c*“ steht für „connected“, deutsch „zusammenhängend“) sind hingegen Ausdruck der Verbundenheit der entsprechenden Molekülbereiche.

Die Variante folgt bei der Backtrackingsuche den *c*-Kanten und sucht nach maximalen Cliquen, deren Knoten alle über einen *c*-Pfad miteinander verbunden sind. Auf diese Weise wird die Suche eingeschränkt. Das Ergebnis ist eine wesentliche Laufzeitverbesserung. Andererseits findet die Variante auf diese Weise lediglich zusammenhängende maximale gemeinsame Subgraphen, die als *c*-MCS bezeichnet werden (vgl. 2.1.4). Für die Betrachtung von Reaktionen ist diese Einschränkung allerdings kein Nachteil. Die Rekonstruktion komplexerer Reaktionen erforderte die Einführung eines Rankingsystems (vgl. 2.2.2), in dem die *c*-MCS nach ihrer Größe sortiert werden. Bei dieser Bewertungsmethode ist es sinnvoller, zusammenhängenden MCS eine höhere Priorität einzuräumen als unzusammenhängenden MCS, die in mehrere Fragmente zersplittert sein können.

Ein effizienter Vergleich von großen Molekülen mit mehr als 60 Atomen ist mit der Variante des Bron-Kerbosch-Algorithmus allerdings noch nicht möglich, auch wenn sie eine deutliche Laufzeitverbesserung mit sich bringt. Ebenso erwies sich der Algorithmus zunächst als zu unflexibel, um komplexere Reaktionen nachzuvollziehen, in denen Ringssysteme gespalten oder gebildet werden. Ursache ist der zugrunde liegende Produktgraph, bei dessen Bildung starke

Einschränkungen vorgenommen werden. Hierdurch fehlen dem Produktgraphen einige Kanten zur Bildung der entsprechenden Cliquen (vgl. 2.1.4, Abbildung 13).

4.1.2.2 McGregor-Algorithmus

Der McGregor-Algorithmus ist ein flexibler MCS-Algorithmus, der für den Molekülvergleich entwickelt wurde. Der Algorithmus benutzt ein Backtracking-Verfahren, um systematisch alle möglichen Zuordnungen der Knoten des einen Graphen auf den anderen Graphen auszuprobieren (vgl. 2.1.5). Allerdings werden bei diesem Verfahren die Bindungen als Knoten betrachtet.
Der McGregor-Algorithmus ist auch in der Lage, die Entstehung oder Spaltung von Ringsystemen nachzuvollziehen. Durch die Integrierung eines „divide and conquer“ (englisch, „teile und herrsche“) Ansatzes in das Backtracking-Verfahren konnte die Laufzeit verbessert werden. Dennoch ist nur ein Vergleich von kleineren Molekülen bis zu einer Größe von etwa 30 Atomen möglich. Neben der Größe hat die Molekülstruktur und Zusammensetzung Einfluss auf die Laufzeit.

4.1.2.3 Kombination der Algorithmen

Der letztlich für den Molekülstrukturvergleich verwendete Algorithmus ist eine Kombination des McGregor-Algorithmus und der Variante des Bron-Kerbosch-Algorithmus. Wie die Variante des Bron-Kerbosch-Algorithmus ist auch die Kombination ein *c*-MCS-Algorithmus. Dies bedeutet, der Algorithmus errechnet die zusammenhängenden maximalen gemeinsamen Substrukturen von zwei Eingabemolekülen (vgl. 2.1.6).
Zu Beginn wird der schnelleren Variante des Bron-Kerbosch-Algorithmus Priorität eingeräumt. Diese errechnet zunächst zusammenhängende gemeinsame Substrukturen. Diese Strukturen werden an den McGregor-Algorithmus übergeben, der überprüft, ob die gemeinsamen Substrukturen bereits maximal sind oder ob sie sich weiter zu maximalen gemeinsamen Substrukturen erweitern lassen. Dabei erweitert der McGregor-Algorithmus die Substrukturen allerdings nur über benachbarte Bindungen. Hierdurch wird sichergestellt, dass weiterhin nur nach zusammenhängenden maximalen gemeinsamen Substrukturen gesucht wird.
Die Kombination der beiden Algorithmen erlaubte auch eine Laufzeitverbesserung für den Vergleich von großen Molekülen. Hierbei werden bei der Bron-Kerbosch-Variante neben dem Elementtyp auch die atomaren Umgebungen bei der MCS-Suche berücksichtigt (vgl. 2.1.6, Abbildung 19). Die gemeinsamen Substrukturen, die durch die Bron-Kerbosch-Variante vorgegeben werden, sind hierdurch zumeist nicht maximal. Erst durch den McGregor-Algorithmus werden die Ränder der Substrukturen ergänzt und somit die *c*-MCS vervollständigt.

Bei kleinen Molekülen ist eine Betrachtung der atomaren Umgebung nicht sinnvoll. Im Verhältnis zur Molekülgröße ist der Anteil der Atome, die an der Reaktion beteiligt sind, höher. Die atomaren Umgebungen innerhalb des Moleküls ändern sich in stärkerem Maße. Die Wahrscheinlichkeit ist daher zu hoch, dass durch Betrachtung der atomaren Umgebungen falsche Zuordnungen entstehen. Außerdem stellt der Vergleich von kleinen Molekülen kein Problem für die Laufzeit dar. Das Programm verzweigt sich daher in Abhängigkeit von der Molekülgröße. Die Umgebungen der Atome werden nur bei dem Vergleich von größeren Molekülen berücksichtigt.

4.1.2.4 Bewertung und Eigenschaften des *c*-MCS-Algorithmus

Durch die Verknüpfung des McGregor-Algorithmus und der Variante des Bron-Kerbosch-Algorithmus ließ sich eine sehr schnelle und flexible *c*-MCS-Variante entwickeln. Die Laufzeitverbesserung erlaubt auch einen Vergleich der größten Moleküle des Datensets.
Bei größeren Molekülen beträgt die *c*-MCS-Suche teilweise noch mehr als eine Minute mit einem Pentium-Rechner (2,4 GHz). Der Vergleich von GM1 (91 Atome ohne Wasserstoffatome) mit GM2 (80 Atome) beträgt beispielsweise 80 Sekunden. Die Laufzeit hängt aber nicht nur von der Molekülgröße ab. Auch die Struktur und Elementzusammensetzung der Moleküle haben großen Einfluss auf die Laufzeit. Die MCS-Suche dauert umso länger, je homogener ein Molekül zusammengesetzt ist. Auch symmetrische und kreisförmige Strukturen innerhalb von Molekülen führen zu einer höheren Rechenzeit. So dauert der Vergleich von Cytochrom mit sich selbst drei mal so lange, wie der Vergleich von GM1 mit GM1, obwohl Cytochrom nur 43 Rückgratatome besitzt und GM1 mit 91 Rückgratatomen mehr als doppelt so groß ist. Cytochrom ist mit seinen vier Pyrrolringen sehr symmetrisch aufgebaut, insbesondere wenn zwischen den Doppel- und Einfachbindungen nicht unterschieden wird. Allerdings bildet Cytochrom unter den mittelgroßen Molekülen eher die Ausnahme. Für die meisten mittelgroßen Moleküle (z.B. FAD und $FADH_2$ mit 53 Atomen) beträgt die Laufzeit deutlich weniger als eine Sekunde.
Die höhere Flexibilität geht auf den McGregor-Algorithmus zurück. Der McGregor-Algorithmus betrachtet die Bindungen als Knoten. Im Backtracking-Suchbaum werden mögliche Bindungszuordnungen als Teillösung angesehen. Bei der Variante des Bron-Kerbosch-Algorithmus wurde hingegen ein Knoten-Produktgraph als Grundlage verwendet. Die Knoten des Produktgraphen werden hierbei aus den Knoten der beiden Eingangsgraphen generiert. Für den Molekülvergleich bedeutet dies, dass die Knoten aus möglichen Atomzuordnungen der beiden Eingangsmoleküle gebildet werden. Der Knoten-Produktgraph wurde als Basis gewählt, da er unmittelbar Aufschluss darüber gibt, welche Atome zu der *c*-MCS gehören. Es wäre allerdings auch

möglich gewesen, einen Kanten-Produktgraphen als Basis zu verwenden, bei dem die Knoten durch Bindungszuordnungen gebildet werden. Wie sich später zeigte, besitzt die Variante des Bron-Kerbosch-Algorithmus eine höhere Flexibilität, wenn sie auf einem Kanten-Produktgraphen basiert. Der McGregor-Algorithmus ist somit nicht generell flexibler als die Varianten des Bron-Kerbosch-Algorithmus. Entscheidend ist vielmehr, ob Atom- oder Bindungszuordnungen betrachtet werden.
Für den Molekülvergleich wurde weiterhin der aus den verschiedenen Algorithmen kombinierte *c*-MCS-Algorithmus verwendet. Entscheidend ist neben der Flexibilität auch der Geschwindigkeitsvorteil, der sich aus der Kombination ergibt.

4.1.3 Atomzuordnung auf Grundlage des Rankingsystems

Der *c*-MCS-Algorithmus besitzt eine zentrale Bedeutung bei der Generierung der Atomzuordnung für enzymatische Reaktionen. Er ermöglicht einen Strukturvergleich der Edukt- und Produktmoleküle. Die Atomzuordnung wird mit Hilfe eines Rankingsystems gebildet. Um auch komplexere Fälle oder Ausnahmen zu berücksichtigen, wurde das System um zusätzliche Verfahren erweitert. Hierzu gehören eine weitere MCS-Variante des Bron-Kerbosch-Algorithmus, ein Algorithmus zum Auffinden aromatischer Ringe, das Entfernen redundanter *c*-MCS, die lokale Symmetrien beinhalten, und eine separate Atomzuordnung der Kofaktoren.

4.1.3.1 Rankingsystem

Der *c*-MCS-Algorithmus ermöglicht einen Strukturvergleich der Edukt- und Produktmoleküle, die paarweise verglichen werden. Da zwei Moleküle meist mehr als eine *c*-MCS besitzen, entsteht bei fast jedem Molekülvergleich ein Set von *c*-MCS. Um Atomzuordnungen für eine Reaktion zu generieren, werden die *c*-MCS aus den Molekülvergleichen miteinander kombiniert. Anschließend werden die *c*-MCS-Kombinationen selektiert, bei denen möglichst viele Atome zugeordnet werden können (vgl. 2.2.1). Wenn nur wenige Moleküle an der Reaktion beteiligt sind, führt häufig allein die Kombination der *c*-MCS zu vollständigen Atomzuordnungen (vgl. Abbildung 21 und 22, Abschnitt 2.2.1).
Häufig sind aber mehr als 3 Moleküle an der Reaktion beteiligt. In diesen Fällen wird eine Atomzuordnung deutlich schwieriger, da es häufig zu Überlagerungen der verschiedenen *c*-MCS innerhalb einer Kombination kommt. Dabei werden einige Atome durch zwei oder mehr *c*-MCS beansprucht. Damit eine eindeutige Abbildung der Atome generiert werden kann, muss eine Entscheidung getroffen werden, welcher *c*-MCS die Atome zugeordnet werden. Bei einer manuellen Zuordnung würden größere *c*-MCS eine höhere Priorität erhalten als kleinere. Daher wurde ein

Rankingsystem entwickelt, das die *c*-MCS innerhalb einer Kombination nach ihrer Größe bewertet (vgl. 2.2.2). Dabei erhalten größere *c*-MCS Vorrang gegenüber den kleineren *c*-MCS einer Kombination. Wenn es Überlagerungen zwischen kleineren und größeren *c*-MCS gibt, werden die entsprechenden Atome den größeren *c*-MCS zugeschrieben. Aus den kleineren *c*-MCS hingegen werden die Atome entfernt.
Bei einigen Reaktionen gibt es mehrere Möglichkeiten, die Eduktatome den Produktatomen zuzuordnen. Dies gilt besonders, wenn symmetrische Strukturen an der Reaktion beteiligt sind oder einzelne Atome übertragen werden, die mehrere Positionen besetzen können (vgl. 2.1.4, Abbildung 13 A). Alle verschiedenen Möglichkeiten werden errechnet und Lösungen mit maximaler Atomzuordnung abgespeichert. Bei den meisten Reaktionen wird durch die *c*-MCS-Berechnung und das Rankingsystem eine vollständige Atomzuordnung erzeugt. Andere Reaktionen erfordern hingegen Erweiterungen des Verfahrens. Hierzu gehören vor allem Reaktionen, bei denen eine hohe Molekülanzahl zu einer explosionsartigen Vervielfachung der Kombinationsmöglichkeiten führt. Außerdem können Probleme bei Reaktionen auftreten, wo kleinere Gruppen übertragen werden, die keiner *c*-MCS angehören.

4.1.3.2 Laufzeitverbesserung durch Entfernung redundanter *c*-MCS, separate Zuordnung der Kofaktoren und Ausschluss überflüssiger Molekülvergleiche

Nach dem Molekülstrukturvergleich ist die Kombination der *c*-MCS in allen möglichen Variationen der zweite laufzeitbestimmende Faktor. In Abhängigkeit von der Anzahl der beteiligten Moleküle und in Abhängigkeit davon, wieviele *c*-MCS bei den einzelnen Molekülvergleichen entstehen, kann die Anzahl der *c*-MCS-Kombinationen schnell sehr große Dimensionen annehmen. Damit die Anzahl von *c*-MCS-Kombinationen nicht zu groß wird, wurden 3 Strategien entwickelt:

1. Entfernung redundanter *c*-MCS:
Die Anzahl der *c*-MCS, die ein Molekülvergleich hervorbringt, variiert sehr stark. Häufig decken aber mehrere *c*-MCS in den Molekülen die gleichen Strukturen ab und sind im Grunde äquivalent. In diesem Fall beruhen die verschiedenen *c*-MCS häufig nur auf lokalen Symmetrien. Zur Erkennung lokaler Symmetrien wurde die *c*-MCS-Suche mit einer Postfilterfunktion verknüpft, die symmetrische Strukturen auffindet und redundante *c*-MCS aus dem Lösungsset entfernt (vgl. 2.2.3).

2. Separate Atomzuordnung der Kofaktoren:

Häufig sind Kofaktoren (vgl. 1.1) an enzymatischen Reaktionen in hoher Anzahl beteiligt. Entsprechend erhöhen sie die Anzahl von *c*-MCS-Kombinationen. Andererseits sind die Veränderungen der Kofaktoren bekannt. Werden auf Edukt- und Produktseite korrespondierende Paare von Kofaktoren gefunden, lassen sich die Moleküle separat zuordnen und stellen keine Belastung bei der *c*-MCS-Kombination dar (vgl. 2.2.4).

3. Ausschluss überflüssiger Molekülvergleiche

Der Vergleich von kleinen Molekülen, wie Methan, Wasser oder Ammoniak, mit großen Molekülen kann zu einer hohen Anzahl von *c*-MCS führen. Denn die kleinen Moleküle finden sich in den Strukturen der größeren Moleküle häufig vielfach wieder. Diese Molekülvergleiche erhöhen die Anzahl von *c*-MCS-Kombinationen wesentlich, obwohl sie zumeist nicht signifikant sind. Es macht daher Sinn, den Vergleich von kleinen mit größeren Molekülen auf eine zweite Molekülvergleichsebene zu verlagern. Kleine Moleküle gehören daher zu den Strukturen, die mit Hilfe des Bron-Kerbosch-Algorithmus zugeordnet werden (vgl. 2.2.6).

4.1.3.3 Bron-Kerbosch-Algorithmus

Das Ranking-System (vgl. 2.2.2) ist bei einigen Reaktionen nicht ausreichend, um die Atomzuordnung abzuschließen. Insbesondere sind Reaktionen problematisch, wo kleine Gruppen von einem Molekül auf ein anderes übertragen werden (vgl. 2.2.5). Wenn diese beiden Moleküle mit Hilfe des *c*-MCS-Algorithmus verglichen werden, kann die errechnete *c*-MCS in einem anderen Bereich der beiden Moleküle liegen. Die kleinen Strukturen, die übertragen werden, können so durch das Rankingsystem nicht zugeordnet werden. Um auch solche Fälle zu berücksichtigen, wurde dem Molekülvergleich eine weitere Ebene hinzugefügt. Diese Vergleichsebene basiert auf einer Variante des Bron-Kerbosch-Algorithmus, die weitgehend dem ursprünglichen Bron-Kerbosch-Algorithmus entspricht und somit auch nicht zusammenhängende MCS findet. Daher kann diese Variante mehrere Strukturen gleichzeitig zuordnen. Der Algorithmus ist auch in der Lage, einzelne Atome zu vergleichen, da er ebenfalls auf einem Knoten-Produktgraphen basiert.

4.1.3.4 Atomzuordnung aromatischer Ringe

Ein weiterer Problemfall bei der Atomzuordnung entsteht durch aromatische Ringe (vgl. 2.2.7). Bei aromatischen Ringen befinden sich 2 Elektronen einer Doppelbindung auf günstigeren Energieniveaus und werden als π-Elektronen bezeichnet. Die π-Elektronen der Ringe sind delokalisiert und verteilen sich in einer Elektronenwolke oberhalb und unterhalb der Ringebene.

In den Moleküldateien, den so genannten Molfiles, werden die Doppelbindungen allerdings auf bestimmte Positionen festgelegt. Die Zuordnung erfolgt dabei willkürlich in Abhängigkeit vom Zeichner. Die Doppelbindungen der aromatischen Ringe können daher in den jeweiligen Edukt- und Produktmolekülen unterschiedlich eingezeichnet sein. Das Ergebnis ist, dass für die Ringe Bindungsumbrüche errechnet werden, obwohl sie an der Reaktion keinen Anteil haben.
Um solche Fehler bei der R-Matrix-Berechnung auszuschließen, wurde ein Algorithmus zur Erkennung von aromatischen Ringen entwickelt. Der Algorithmus basiert auf einem Backtrackingverfahren. Im Wesentlichen erfolgt die Suche auf der Grundlage der Aromatizitätskriterien. Der Algorithmus ermöglicht es, die aromatischen Ringe der Eduktseite und Produktseite einander zuzuordnen. Ist ein Ring vor als auch nach der Reaktion aromatisch, können keine Bindungsumbrüche innerhalb des Ringes aufgetreten sein. Mit Hilfe des Algorithmus können solche Fälle erkannt und ausgeschlossen werden.

4.1.3.5 Generierung und Zuordnung der Wasserstoffatome

Das Addieren und die Zuordnung der Wasserstoffatome schließt die Atomzuordnung ab (vgl. 2.2.8). Die Molfiles der Moleküle enthalten zumeist keine Wasserstoffatome, bis auf wenige Ausnahmen, wie elementarer Wasserstoff. Es sind vorwiegend nur die schwereren Atome des Molekülrückgrates in den Molefiles kodiert. Das Molekülrückgrat besteht aus Kohlenstoff- oder Heteroatomen, wie beispielsweise Sauerstoff, Stickstoff oder Schwefel. Im Gegensatz zu diesen Atomen können Wasserstoffatome nur eine Bindung eingehen. Sie lassen sich daher bestimmten Atomen des Molekülrückgrates zuordnen. Die Anzahl der Wasserstoffatome eines Rückgratatomes hängt ab von seinem Elementtyp, seiner Ladung und der Anzahl seiner Bindungen zu den anderen Atomen des Molekülrückgrats. Ebenso wie die Rückgratatome erhalten die neu erzeugten Wasserstoffatome eine eindeutige Nummer, um sie von allen anderen Atomen des Reaktionssets eindeutig unterscheiden zu können.
Wenn die Generierung der Wasserstoffatome abgeschlossen ist, werden die Eduktwasserstoffatome den Produktwasserstoffatomen zugeordnet. Die Grundlage hierfür ist die Atomzuordnung des Molekülrückgrates, die bereits durch das Rankingsystem errechnet wurde. Dabei werden die Eduktwasserstoffatome eines Rückgratatoms seinen Produktwasserstoffatomen zugeordnet. Die Wasserstoffatome werden auf zweidimensionaler Ebene betrachtet und unterscheiden sich nicht in ihren Eigenschaften. Daher werden die Wasserstoffatome eines Rückgratatoms der Nummerierung folgend willkürlich zugeordnet.

Die Anzahl der Wasserstoffatome eines Rückgratatoms vor und nach der Reaktion wird verglichen. So kann festgestellt werden, ob ein Rückgratatom ein Wasserstoffatom verloren oder hinzugewonnen hat. Auf diese Weise werden die abgespaltenen Wasserstoffatome der Eduktseite und die neu gebundenen Wasserstoffatome der Produktseite identifiziert.

Da die Wasserstoffatome, die auf der Eduktseite abgespalten werden, formal zu den Wasserstoffatomen gehören müssen, die auf der Produktseite neue Bindungen eingehen, werden diese Atome einander zugeordnet. Die Zuordnung erfolgt auch hier willkürlich, denn es lässt sich in der Regel schwer nachvollziehen, welche Position ein abgespaltenes Wasserstoffatom auf der Produktseite besetzt - zumal sich die Moleküle in wässriger Lösung befinden, die H^+-Ionen enthält, die in die Reaktion involviert werden können.

Bei vielen Reaktionen wurde mit diesem Schritt eine 1:1 Zuordnung der Atome erreicht. Dies betrifft allerdings nicht Reaktionen, in die H^+-Ionen eingehen oder freigesetzt werden. In diesem Fall stimmt die Anzahl der Wasserstoffatome auf der Eduktseite und Produktseite nicht überein. Durch einen Abgleich der Wasserstoffatome der Rückgratatome werden schnell die Stellen identifiziert, wo die H^+-Ionen abgespalten werden oder hinzukommen. Auf der jeweils gegenüber liegenden Seite wird dann ein H^+-Ion hinzu addiert. H^+-Ionen erhalten alle die Atomnummerierung „-101“, um sie von den anderen regulären Atomen unterscheiden zu können.

4.1.4 Berechnung der R-Matrizen

Wenn eine Zuordnung der Eduktatome auf die Produktatome vorliegt, ist die R-Matrix-Berechnung vergleichsweise einfach (vgl. 2.3). Die Atomzuordnung ermöglicht es, die Eduktmatrix (B) und die Produktmatrix (E) so zu konstruieren, dass die korrespondieren Edukt- und Produktatome den gleichen Index erhalten. Die Berechnung der R-Matrix erfolgt anhand von Gleichung 2 (vgl. 1.3.2). Die R-Matrix zeigt, wie die Elektronen zwischen den Atomen während einer Reaktion ausgetauscht werden. Sie gibt Aufschluss über die Spaltung und Entstehung von Bindungen oder darüber, ob sich die Valenzelektronenzahl eines Atoms verändert.

Die Matrizen B, E und R enthalten zunächst sämtliche Atome von allen Molekülen des Reaktionssets. Da häufig größere Moleküle in hoher Anzahl an einer Reaktion beteiligt sind, entstehen leicht Matrizen mit einer Dimension von mehreren hundert Zeilen und Spalten. Erst durch die Berechnung der R-Matrix wird ersichtlich, welche Atome in die Reaktion involviert sind. Da zumeist nur wenige Atome an der Reaktion beteiligt sind, bestehen die R-Matrizen überwiegend aus Zeilen und Spalten, die nur aus Null-Einträgen bestehen (vgl. 2.3, Abbildung 33). Daher werden die Matrizen auf die Atome reduziert, die an der Reaktion beteiligt sind. Optional kann auch nur die R-

Matrix zusammen mit den zugehörigen Edukt- und Produktatomen ausgegeben werden. Diese Art der Darstellung erwies sich für den Gebrauch am übersichtlichsten (vgl. 2.3, Abbildung 34).
Die Darstellung von einigen Isomerasereaktionen erforderte besondere Methoden (vgl. 2.3.1). Die Reaktionen von einigen Isomerasen, wie die der intramolekularen Oxidoreduktasen oder der intramolekularen Transferasen, lassen sich zwar mit dem herkömmlichen Dugundji-Ugi-Modell beschreiben. Anders verhält es sich mit den Isomerasereaktionen, die Veränderungen der absoluten Konfiguration katalysieren. Veränderungen in der Stereoisomerie lassen sich nicht über Bindungsumbrüche oder eine Änderung der Valenzelektronenzahl darstellen. Daher würde die R-Matrix nur Nulleinträge beinhalten.
Abweichend vom ursprünglichen Modell wurde deshalb die absolute Konfiguration in das System aufgenommen. Danach wird eine S Konfiguration durch eine „-10“ und eine R Konfiguration durch eine „10“ auf der Hauptdiagonalen der B und E Matrizen gekennzeichnet. Änderungen in der absoluten Konfiguration werden durch Werte, wie 20 oder -20, in der R-Matrix ersichtlich (vgl. 2.3.1, Abbildung 36).

4.1.5 Kanonisierung der R-Matrizen

4.1.5.1 R-Matrix-Vergleichsproblem

Schwieriger als die R-Matrix-Berechnung ist der Vergleich der R-Matrizen. Von jeder R-Matrix lassen sich n! mögliche Permutationen erzeugen, da die Reihenfolge der Atome beliebig verändert werden kann. Um demnach zwei R-Matrizen auf Identität zu überprüfen, müssten theoretisch alle n! Permutationen beider R-Matrizen erzeugt und miteinander verglichen werden. Bei einer Matrixdimension von 7 existieren aber bereits 5040 Permutationen und mehr als 25 Millionen Vergleiche sind maximal erforderlich, um beide Matrizen zu vergleichen. Da bei den biochemischen Reaktionen des Datensets die Matrixdimension zwischen 1 und 90 variiert, ist eine solche Methode nicht anwendbar.
Eine Lösung für dieses Problem bietet das Verfahren der Kanonisierung (vgl. 2.4). Es erlaubt, aus der Vielzahl von Permutationen eine repräsentative R-Matrix zu selektieren. Diese so genannte kanonische R-Matrix enthält eine kontinuierliche Folge von positiven und negativen Einträgen entlang der ersten Nebendiagonalen. Diese Einträge sollen in den Vorzeichen alternieren und mit einem negativen Eintrag beginnen. Diese Regeln ergeben sich aus den Analysen von Elektronentransfermustern, die häufig eine zyklische Struktur besitzen. Dies wird besonders offensichtlich, wenn die Atome verkettet werden, die miteinander durch Bindungsumbrüche in Wechselbeziehung stehen.

Diese Wechselbeziehungen der Atome lassen sich gut darstellen, wenn die R-Matrix in einen R-Graphen übertragen wird, wobei die Atome die Knoten und die Matrixeinträge die Kanten darstellen können (vgl. 2.4.1). Der längste Pfad innerhalb des R-Graphen gibt die Reihenfolge der Atome in der kanonischen R-Matrix an.

Der R-Graph ist häufig linear oder zyklisch aufgebaut. In diesem Fall lässt sich für die erste Nebendiagonale immer eine fortlaufende Folge von Einträgen mit Vorzeichenwechsel konstruieren. Auf dieser Annahme basieren ältere Kanonisierungsalgorithmen (Brandt *et al.*, 1981, Brandt *et al.*, 1983). Die R-Graphen biochemischer Reaktionen erwiesen sich jedoch als komplexer. So können die R-Graphen Verzweigungen aufweisen und das Elektronentransfermuster kann unterbrochen werden, wodurch mehrere Graphfragmente entstehen.

Verzweigungen entstehen dann, wenn ein Atom auf der Eduktseite die Bindung zu mehreren Atomen auflöst oder auf der Produktseite mit mehreren Atomen eine neue Bindung eingeht. Ein nicht zusammenhängender Graph aus mehreren Fragmenten entsteht, wenn Atome ausschließlich an der Spaltung einer Bindung oder an der Entstehung einer Bindung beteiligt sind. Hierzu zählen Metallatome, andere Atome des Molekülrückgrats und vor allem H^+-Ionen (vgl. 2.4.2).

Metallatome, wie Eisen-, Kupfer- oder Cobaltatome, spielen als Reduktions- oder Oxidationsmittel bei biochemischen Reaktionen eine Rolle. Die aufgenommenen oder abgegebenen Elektronen verändern allerdings nur die Valenzelektronenzahl der einzelnen Metallatome. Ebenso beziehen sich Veränderungen in der absoluten Konfiguration nur auf einzelne Atome. Daher sind diese Atome innerhalb des R-Graphen isoliert. Hinzu kommen Atome des Molekülrückgrates, die beispielsweise den Verlust einer Bindung lediglich durch eine Änderung der Valenzelektronenzahl kompensieren.

Am häufigsten führen H^+-Ionen zu einer Fragmentierung des R-Graphen. H^+-Ionen können in Reaktionen eingehen und werden dort an Moleküle gebunden oder werden von Molekülen abgespalten und freigesetzt. Die H^+-Ionen sind deshalb entweder nur an der Spaltung oder Bildung einer Bindung beteiligt. In wieviele Fragmente ein R-Graph gespalten wird, hängt zum einen von der Anzahl der H^+-Ionen ab. Wenn nur ein oder zwei H^+-Ionen an einer Reaktion beteiligt sind, wird ein zyklischer R-Graph lediglich linearisiert. Spätestens, wenn 3 H^+-Ionen an der Reaktion beteiligt sind, ist die Fragmentierung in einen nicht zusammenhängenden R-Graph unvermeidlich (vgl. 2.4.2, Abbildung 40).

4.1.5.2 Kanonisierungsalgorithmus

Die unter Abschnitt 4.4.1 beschriebenen Problemfälle machen die Notwendigkeit deutlich, warum ein neuer Algorithmus für die Kanonisierung von R-Matrizen biochemischer Reaktionen entwickelt

wurde. Der Kanonisierungsalgorithmus basiert auf einem Backtracking Verfahren, das alle längsten Pfade des R-Graphen errechnet. Wenn der R-Graph nicht zusammenhängend ist, werden alle längsten Pfade innerhalb der verschiedenen Graphfragmente errechnet. Die längsten Pfade der Graphfragmente lassen sich in verschiedenen Reihenfolgen und in verschiedenen Orientierungen verknüpfen. Hierdurch können aus einer R-Matrix leicht einige hundert verkettete Pfade konstruiert werden. Aus den verketteten Pfaden lassen sich die Atomreihenfolgen ableiten, die potentielle Kandidaten für die Konstruktion von kanonischen Matrizen darstellen. Um das Set von Atomreihenfolgen zu verringern, wird die Anzahl der verketteten Pfade mit Hilfe der Kanonisierungskriterien verringert.

So werden nach Möglichkeit Pfade selektiert, die mit einer negativen Kante beginnen. Außerdem sollten die Kanten eines Pfades in ihren Vorzeichen alternieren. Manchmal ist dieses aber nicht möglich, wenn beispielsweise 2 positive Kanten aufeinander folgen oder wenn es sich um einen verketteten Pfad aus mehreren Fragmenten handelt. Ein weiterer wichtiger Selektionsfaktor ist das Elementgewicht der Atome. Dabei wird Pfaden Vorrang eingeräumt, bei denen die schwersten Elemente eine vordere Position innerhalb des Pfades besetzen. Einen Überblick über das Kanonisierungsschema gibt Abbildung 41 (vgl. 2.4.2). Auf diese Weise gelingt es, für die meisten Reaktionen nur einen Pfad zu selektieren, aus dem eine kanonische Matrix aufgebaut werden kann. Nur in wenigen Fällen entstehen 2, 3 oder 4 kanonische Matrizen.

Neben dem paarweisen Molekülvergleich durch den *c*-MCS-Algorithmus und der Generierung der Atomzuordnung durch das Rankingsystem bildet das Kanonisierungsverfahren die dritte laufzeitbestimmende Komponente des Programms. Der wichtigste Bestandteil der Kanonisierung ist der Backtracking-Algorithmus, der alle längsten Pfade innerhalb des R-Graphen findet. Backtracking Algorithmen stellen eine elegante Lösung für komplexe Probleme dar und werden häufig in der Graphentheorie verwendet (vgl. 2.1.3). Allerdings führen komplexe Graphen schnell zu sehr hohen Laufzeiten. R-Graphen sind allerdings meist einfach aufgebaut. Sie besitzen häufig eine lineare Struktur und beinhalten nur wenige Verzweigungen. Ein größeres Problem stellen hingegen nicht zusammenhängende R-Graphen dar, die aus vielen Graphfragmenten gleicher Größe bestehen. Hier ergeben sich häufig viele Möglichkeiten, die längsten Pfade der Graphfragmente miteinander zu verknüpfen. Eine Rolle für die Laufzeit spielt auch die Zuordnung der Wasserstoffatome, wenn H^+-Ionen an der Reaktion beteiligt sind. Je nachdem welchen Wasserstoffatomen die H^+-Ionen zugeordnet werden, kann der R-Graph mehr oder weniger fragmentiert sein (vgl. 2.4.2). Um die R-Graphen zu finden, die am wenigsten fragmentiert sind, müssen daher alle möglichen Wasserstoffatomzuordnungen konstruiert und analysiert werden. An der Nitrogenase Reaktion sind allerdings 24 Wasserstoffatome beteiligt, womit 24! Möglichkeiten

existieren, die Wasserstoffatome einander zuzuordnen. Da dies nicht umsetzbar ist, wurde auf diese Analyse verzichtet, wenn mehr als 6 Wasserstoffatome an der Reaktion beteiligt sind.

4.1.5.3 Reaktionsvergleich auf der Basis von R-Strings

Um den Vergleich der kanonisierten R-Matrizen zu erleichtern, werden sie in eine String-Notation überführt (vgl. 2.4.3). Je R-Matrix werden 2 R-Strings generiert. Der erste ist ein Integer-String und setzt sich aus den Matrixeinträgen zusammen. Der zweite String, der Char-String, wird aus den Elementsymbolen der beteiligten Atome gebildet. Der Char-String bezieht sich auf die Einträge des Integer-Strings. Er zeigt an, welche Atome mit dem jeweiligen Matrixeintrag in Verbindung stehen.

4.2 R-Matrix-Berechnung biochemischer Reaktionen

Auf der Basis eines größeren Datensets, das fast alle Subsubklassen des EC-Systems abdeckt, wurden die R-Matrizen von mehreren tausend biochemischen Reaktionen berechnet. Hiermit war es möglich, die Elektronentransfermuster der Enzyme zu vergleichen und zu neuen Gruppen zusammenzufassen. Da das Dugundji-Ugi-Modell ebenso wie das EC-Klassifikationssystem auf den chemischen Eigenschaften der Gesamtreaktion beruht, waren die Unterschiede und Gemeinsamkeiten der beiden Systeme von besonderem Interesse.

4.2.1 Das Datenset

Die enzymatischen Reaktionen und die zugehörigen Moleküle stammen aus der Braunschweiger Enzymdatenbank BRENDA (Schomburg 2004). Das Datenset enthält 3330 enzymatische Reaktionen und 3876 Moleküle, die in Textdateien, den Molfiles, abgespeichert sind. Hiermit enthält das Datenset weniger als die etwa 4000 bekannten Enzyme. Allerdings wurde das Datenset in den letzten Jahren ständig erweitert. Heute deckt es 228 Subsubklassen der 229 bislang formulierten Subsubklassen ab. Die Abdeckung der Subsubklassen war bei der Wahl des Datensets besonders wichtig, da erwartet werden konnte, dass die Elektronentransfermuster innerhalb einer Subsubklasse bereits sehr homogen sind. Das EC-Klassifikationsystem sieht vor, dass die Reaktionen dem EC-Nummerncode folgend immer spezifischer eingeteilt werden. Auf der Ebene der Subsubklassen sind die Reaktionen bereits soweit spezifiziert, dass die beteiligten Bindungen,

Elektronenakzeptoren, Kofaktoren und stereochemischen Veränderungen meist identisch sind. Dem Datenset fehlt lediglich eine Subsubklasse. Diese Subsubklasse gehört zur Subklasse der Intramolekularen Oxidoreduktasen. Es handelt sich um die Subsubklasse 5.3.4.-, die nur ein Enzym, die Protein Disulfid-Isomerase (EC-Nummer 5.3.4.1), enthält. Diese katalysiert die Verlagerung von Disulfid-Bindungen in Proteinen.
Die Moleküle der BRENDA-Datenbank wurden so konstruiert, dass Veränderungen an den Molekülen automatisch nachvollzogen werden können. Dies ist ein weiterer wichtiger Vorteil des BRENDA-Datensets. Zu Beginn der vorliegenden Arbeit wurden das Reaktionsset und die Molfiles der KEGG (Kyoto Encyclopedia of Genes and Genomes)-Datenbank verwendet. Die KEGG-Datenbank (Kanehisa, 2008) verfügt ebenfalls über ein umfangreiches Set von enzymatischen Reaktionen (~ 4300). Die Reaktionen des Datensets lassen sich aber nicht immer automatisch nachvollziehen. Beispielsweise werden langkettige Polymere auf der Edukt- und Produktseite durch identische Moleküle beschrieben.

4.2.2 Problemreaktionen

Von den 3330 zur Verfügung stehenden enzymatischen Reaktionen des Datensets wurden die R-Matrizen von 3209 Reaktionen automatisch errechnet. Die Ursachen hierfür sind im Einzelnen vielfältig. Im Wesentlichen handelt es sich um Laufzeitprobleme, die besonders bei zwei Verfahren im Programm auftreten können.
Der erste Problembereich ist die *c*-MCS-Kombination im Zusammenspiel mit dem Rankingsystem. Wenn viele Edukt- und Produktmoleküle an der Reaktion beteiligt sind, können einige Tausend *c*-MCS-Kombinationen (vgl. 2.2.2) entstehen. Besonders problematisch kann die Situation werden, wenn es außerdem noch innerhalb einer *c*-MCS-Kombination einige *c*-MCS gleicher Größe gibt. Das Ranking-System ordnet die *c*-MCS innerhalb einer Kombination nach ihrer Größe, wobei größere *c*-MCS Vorrang vor kleineren *c*-MCS erhalten. Gibt es aber innerhalb einer Kombination mehrere *c*-MCS, so kann zunächst keiner *c*-MCS Vorrang eingeräumt werden. In diesem Fall werden alle Möglichkeiten ausgetestet, die gleich großen *c*-MCS nach Priorität anzuordnen. Durch diesen Schritt kann sich die Anzahl der *c*-MCS-Kombination explosionsartig vervielfachen.
Damit die Anzahl der *c*-MCS-Kombination nicht zu groß wird, wurden bereits heuristische Aspekte in das Verfahren integriert (vgl. 4.1.3.2). So werden bei dem paarweisen Molekülvergleich durch den *c*-MCS-Algorithmus lokale Symmetrien erkannt und redundante aus dem Lösungsset entfernt (vgl. 2.2.3). Die Atomzuordnung der Kofaktoren erfolgt getrennt von den übrigen Molekülen der Reaktion (vgl. 2.2.4). Durch das Hinzufügen einer weiteren Bron-Kerbosch-Variante wurde eine weitere

Molekülvergleichsebene geschaffen, die es erlaubt, kleine Moleküle zunächst der Atomzuordnung zu entziehen (vgl. 2.2.6). Die Möglichkeiten, an dieser Stelle weitere Verbesserungen vorzunehmen, sind aber noch nicht ausgeschöpft. So bringen Erweiterungen des Datensets neue Kofaktoren mit sich, die in das System integriert werden müssen. Eine andere Möglichkeit bestünde in einer grundlegenden Überarbeitung des *c*-MCS-Kombinationssystems und des Rankingsystems. Gegenwärtig werden alle Möglichkeiten der Atomzuordnung errechnet. Um die Laufzeit zu verkürzen, wäre es günstiger, bei problematischen Fällen die Suche vorzeitig abzubrechen und sich auf erste Lösungen zu beschränken. Dies ist gegenwärtig nicht möglich, da das Verfahren wie eine Breitensuche angelegt ist.

Der zweite Problembereich des Verfahrens, der zu einem Programmabbruch führen kann, ist die Kanonisierung (vgl. 2.4). Das Hauptproblem sind Reaktionen mit fragmentierten Elektronenaustauschmustern. Hierbei ergeben sich häufig viele Möglichkeiten, die längsten Pfade der Graphfragmente miteinander zu verknüpfen. Besonders wenn einige Graphfragmente die gleiche Größe haben, werden alle möglichen Permutationen getestet. Reaktionen, bei denen eine hohe Anzahl von H^+-Ionen beteiligt sind, oder Isomerasereaktionen gehören zu dieser Problemklasse (vgl. 4.1.5). Um die Laufzeitprobleme zu umgehen, wurden auch hier heuristische Verfahren in den Kanonisierungsmechanismus integriert. Dennoch sind noch weitere Verbesserungsmöglichkeiten vorstellbar.

4.2.3 Matrixdimensionen der Subsubklassen

Die Matrixdimension ist ein wichtiger Kennwert, der anzeigt, wie viele Atome an der Reaktion beteiligt sind. Die Matrixdimensionen biochemischer Reaktionen variieren zwischen 1 und 90 (vgl. 3.2). Eine Matrixdimension von 1 besitzen die Isomerasereaktionen der Subklasse 5.1.-.-. Bei den Enzymen dieser Subklasse handelt es sich um Racemasen, die Veränderungen in der Stereo-Isomerie katalysieren. Dabei verändert sich die absolute Konfiguration eines Atoms. Die Matrixdimensionen 2 und 3 werden nur durch Oxidoreduktasen besetzt.

Dagegen ist die Matrixdimension von 4, mit 93 Subsubklassen, weitaus am häufigsten vertreten. Die Subsubklassen mit der Matrixdimension 4 stammen aus den ersten 4 Hauptklassen des EC-Klassifikationssystems. Hierzu gehören einige Subsubklassen aus der Hauptklasse der Oxidoreduktasen.

Die Matrixdimension 4 ist aber am weitesten verbreitet unter den Subsubklassen der Transferasen, Hydrolasen und Lyasen. Bei vielen Reaktionen dieser Hauptklassen werden 2 Bindungen auf der Eduktseite gespalten, während auf der Produktseite 2 Bindungen gebildet werden. Dieses

Reaktionsmuster ist typisch für viele biochemische Reaktionen. So übertragen Transferasen Molekülstrukturen von einem Molekül auf ein anderes. Hierfür müssen auf der Eduktseite 2 Bindungen gespalten werden. Auf der Produktseite entstehen hingegen 2 Bindungen. Die Hydrolasen und Lyasen, die Moleküle aufspalten, besitzen ebenso dieses Reaktionsmuster.
In der Regel sind an der Spaltung und Entstehung von Bindungen je 2 Atome beteiligt. Daher besitzt der überwiegende Teil der enzymatischen Reaktionen (190 Subsubklassen) eine gerade R-Matrix Dimension. Ungerade R-Matrixdimensionen findet man dagegen überwiegend nur bei den Oxidoreduktasen und Isomerasen. Es gibt verschiedene Ursachen für ungerade Matrixdimensionen. Sie entstehen beispielsweise, wenn ein Atom an der Entstehung oder Spaltung von mehreren Bindungen zu mehreren Atomen beteiligt ist (vgl. 3.3, R-Matrix der Violaxanthin de-Epoxidase, Abbildung 47). Eine andere Möglichkeit kann darin bestehen, dass ein Atom eine Änderung seiner Bindungsanzahl durch ein freies Elektronenpaar ausgleicht (vgl. 3.3, R-Matrix der Phosphoadenylyl-sulfat-Reduktase, Abbildung 51). Auch Veränderungen, die sich nur auf ein Atom auswirken, können zu ungeraden R-Matrixdimensionen führen. Hierzu gehören beispielsweise Isomerasereaktionen, die Änderungen in der absoluten Konfiguration katalysieren, oder Oxido-reduktasereaktionen, die Änderungen in der Valenzelektronenzahl von Metallatomen beinhalten.
Die Subsubklassen der Oxidoreduktasen sind in fast allen vertretenen R-Matrixdimensionen zu finden. An Oxidoreduktasereaktionen sind häufig Kofaktoren beteiligt, wie NAD, FAD oder ATP. Diese Kofaktoren können in geringer oder hoher Anzahl an Oxidoreduktasereaktionen beteiligt sein und haben so maßgeblich Einfluss auf die Matrixdimension. Die Subsubklassen der ersten Hauptklasse bilden daher die Gruppe mit den höchsten R-Matrixdimensionen. Aus diesem Grund sind ab einer R-Matrixdimension von 10 nur noch Oxidoreduktasereaktionen zu finden. Die Reaktion der Nitrogenase verfügt bei einer Dimension von 90 über die größte R-Matrix. An der Reaktion sind über 40 Moleküle beteiligt.

4.2.4 Homogenität der Elektronentransfermuster innerhalb der Subsubklassen

Eine der besten Möglichkeiten, das Dugundji-Ugi-Modell mit den EC-Klassifikationssystem zu vergleichen, bietet eine Analyse der Subsubklassen des EC-Systems. Die EC-Klassifizierung wird ausgehend von den Hauptklassen über die Subklassen bis hin zu den Subsubklassen immer spezifischer. Auf der Ebene der Subsubklassen sind die Reaktionen schon so weit spezifiziert, dass sie weitgehend identisch sind in Hinblick auf die beteiligten Bindungen, Kofaktoren oder stereochemischen Veränderungen. Aus diesem Grund müssten auch die Elektronentransfermuster innerhalb der Subsubklassen einheitlich sein. Stellvertretend für die kanonischen Matrizen wurden

die R-Strings innerhalb einer Subsubklasse verglichen und die R-Stringidentität errechnet. Als R-Stringidentität wurde dabei der prozentuale Anteil des häufigsten R-Strings an den Reaktionen der Subsubklasse definiert.

Die Elektronentransfermuster innerhalb der Subsubklassen erwiesen sich mit einer durchschnittlichen Übereinstimmung von 76 Prozent als homogen. Mehr als 59 Prozent der Subsubklassen besitzen eine R-Stringidentität zwischen 75 und 100 Prozent. Bei 82 Prozent der Subsubklassen liegt die R-Stringidentität über 50 Prozent. Allerdings variiert die Anzahl der EC-Nummern bei den Subsubklassen sehr. Sechzig Subsubklassen sind nur durch die Reaktion einer EC-Nummer vertreten. Bei diesen Subsubklassen liegt die R-Stringidentität somit zwangsläufig bei 100 Prozent. Wenn die Anzahl der EC-Nummern einer Subsubklasse als Gewichtung herangezogen wird, beträgt die R-Stringidentität aber immerhin noch 69,45 Prozent.

Bei 18 Prozent der Subsubklassen liegt die R-Stringidentität unterhalb von 50 % (vgl. 3.3, Abbildung 44). Hierzu gehören unter anderem die Subsubklassen des Typs a.b.98.- oder a.b.99.-. In diesen Subsubklassen werden Reaktionen zusammengefasst, die in keine andere Subsubklasse der Subklasse passen. Die unter 3.3 betrachtete Subsubklasse 1.10.99.- ist ein gutes Beispiel hierfür. Sie gehört der Subklasse 1.10.-.- an, die in 4 Subsubklassen unterteilt ist. Die Enzyme dieser Subklasse haben gemeinsam, dass Diphenole oder verwandte Substanzen als Elektronendonatoren fungieren. Die Subsubklasseneinteilung erfolgt auf Grundlage der verwendeten Elektronenakzeptoren. Bei der ersten Subsubklasse dient NAD+/NADP+, bei der zweiten Cytochrom und bei der dritten Sauerstoff als Elektronenakzeptor. In der vierten Subsubklasse 1.10.99.- werden hingegen die Reaktionen der Subklasse zusammengefasst, bei denen der Elektronenakzeptor nicht NAD+/NADP+, Cytochrom oder Sauerstoff ist. Die Zuordnung erfolgt nach dem Ausschlussprinzip. Entsprechend heterogen sind die drei Reaktionen dieser Subsubklasse, bei denen keine der R-Matrizen übereinstimmt (vgl. 3.3, Abbildung 45 - 47).

Die niedrigste R-Stringidentität besitzen die Subsubklassen der 5. Hauptklasse. Im Gegensatz zu den anderen Hauptklassen wird bei den Isomerasereaktionen die dreidimensionale Struktur berücksichtigt. Insbesondere Änderungen in der absoluten Konfiguration (vgl. 2.3.1) gehen in die R-Matrizen ein. Die dreidimensionale Struktur eines Moleküls hängt vom jeweiligen Zeichner des Moleküls ab. Das Zeichenprogramm MDL ISIS Draw ist in der Lage, die absolute Konfiguration eines Atoms anhand der Molekülzeichnung automatisch zu errechnen und entsprechend im Molfile zu kennzeichnen. Für die Berechnung der absoluten Konfiguration wird berücksichtigt, ob Bindungen aus der Molekülebene hervorstehen oder hinter die Ebene zurücktreten. Hierbei können in Abhängigkeit vom Fachwissen des jeweiligen Zeichners Unterschiede bei korrespondierenden Edukt- und Produktmolekülen auftreten, die über die eigentlichen Veränderungen hinausgehen. Das

Ergebnis sind unterschiedliche R-Matrizen, die sich nicht gruppieren lassen. Aus diesem Grund wurden Veränderungen in der absoluten Konfiguration nur bei der 5. Hauptklasse berücksichtigt.
Neben diesen beiden besonderen Gruppen von Subsubklassen gibt es weitere Subsubklassen, die im Detail genauer studiert werden müssen, um die geringe Homogenität der R-Matrizen nachvollziehen zu können. Ein solcher Fall ist die Subsubklasse 1.8.4.-, die aus kleineren und größeren Gruppen mit identischen R-Strings besteht. Von dieser Subsubklasse sind 11 Reaktionen in dem Datenset enthalten. Bei fast allen Reaktionen werden Disulfidbindungen gespalten oder gebildet. Allerdings zeigen sich große Abweichungen bei den übrigen Bindungen, die an der Reaktion beteiligt sind. Daher gliedert sich die Subsubklasse in 6 Gruppen, wenn die Reaktionen nach ihren R-Strings gruppiert werden.
Die größte Gruppe enthält 3 Reaktionen, womit die R-Stringidentität lediglich bei 27,27 Prozent liegt. Bei allen Reaktionen dieser Gruppe kommt es zu einer Parallelreaktion, wobei sich ein Sulfitmolekül an eine Phospatgruppe anlagert. Aus dieser Parallelreaktion stammen die Elektronen, die eine Spaltung der Disulfidbindung ermöglichen. Bei den anderen Gruppen der Subsubklasse können die Elektronen aber auch aus S-H-Bindungen, Wassermolekülen oder anderen Donatoren stammen. Bei zwei Reaktionen wird das Sauerstoffatom einer S=O Bindung zu Wasser reduziert, wobei eine Disulfidbindung gespalten wird. Es zeigt sich, dass sich die Reaktionen im Detail doch sehr unterscheiden, auch wenn bei allen Reaktionen Disulfidbindungen beteiligt sind.
Neben der Subsubklasse 1.8.4.- gibt es zahlreiche weitere Subsubklassen mit sehr heterogenen Elektronentransfermustern, die im Ergebnisteil aufgeführt sind. Die Unterschiede resultieren meist aus Parallelreaktionen von Kofaktoren oder anderen Molekülen. Wenn diese Parallelreaktionen nicht auf Kofaktoren zurückgehen, lassen sie sich automatisch schwer erkennen und von der eigentlich relevanten Reaktion trennen. Andererseits sind auch die Parallelreaktionen Bestandteil der Gesamtreaktion und machen einen Teil der Enzymspezifität aus.

4.2.5 Gruppierung der Subsubklassen nach R-String Identität

Inwiefern das Dugundji-Ugi-Modell dazu beitragen kann, die EC-Klassifizierung zu unterstützen, zeigt am besten eine neue Einteilung der enzymatischen Reaktionen anhand ihrer Elektronentransfermuster.
Um einen übersichtlichen Vergleich des Dugundji-Ugi-Modells gegenüber dem EC-Klassifikations-systems zu ermöglichen, wurden wieder die Subsubklassen als Basis verwendet. Hierfür wurden die R-Strings verglichen und für jede Subsubklasse der häufigste R-Sting errechnet. Dieser R-String wurde als repräsentativ für die jeweilige Subsubklasse angesehen. Schließlich wurden die R-Strings

der verschiedenen Subsubklassen verglichen und Subsubklassen mit identischen R-Strings gruppiert (vgl 3.4).

121 von 228 Subsubklassen besitzen ein spezifisches Elektronentransfermuster, das sich nicht gruppieren lässt. Theoretisch ließe sich bei den Reaktionen aus dieser Gruppe die Subsubklasse anhand ihres R-Strings mit hoher Wahrscheinlichkeit vorhersagen, vorausgesetzt die Subsubklasse ist sehr homogen.

107 Subsubklassen besitzen hingegen kein spezifisches Elektronentransfermuster. Sie bilden 26 Gruppen verschiedener Größe (vgl 3.4, Tabelle 5). Bei den verschiedenen Subsubklassen einer Gruppe werden die gleichen Bindungen gespalten und gebildet. Zudem kann sich die Valenzelektronenzahl eines Atoms gleichen Elementtyps um den gleichen Wert verändern, oder es erfolgt eine identische Veränderung der absoluten Konfiguration.

Die größten Gruppen enthalten Subsubklassen aus den Hauptklassen der Transferasen, Hydrolasen und Lyasen. Wie in Tabelle 4 (vgl. 3.2) beschrieben wird, besitzen die Subsubklassen dieser 3 Hauptklassen fast alle eine R-Matrixdimension von 4. Daher verwundert es nicht, dass auch die R-Matrizen der größten Gruppen ebenfalls eine Dimension von 4 haben. Diese Matrixdimension ist typisch, wenn 2 Bindungen auf der Eduktseite gespalten werden und 2 Bindungen auf der Produktseite entstehen. Dies ist ein weit verbreitetes Muster unter biochemischen Reaktionen (vgl. 3.2 und 4.2.3). Daher besitzen immerhin 13 der 26 Gruppen eine R-Matrixdimension von 4. Eine Analyse der größten Gruppen zeigt, dass die chemischen Eigenschaften der Molekülstrukturen, die am Reaktionskern beteiligt sind, weitgehend identisch sind.

Bei allen Reaktionen der größten Gruppe wird auf der Eduktseite eine C-O und O-H Bindung gespalten, während auf der Produktseite eine C-O und O-H Bindung entsteht. Diese Bindungstypen treten häufig in Molekülen auf und sind oft an Reaktionen beteilig. Daran haben vor allem die Sauerstoffatome Anteil mit ihren freien Elektronenpaaren und dem elektronenziehenden Effekt auf Bindungen. So zum Beispiel bei Hydrolasereaktionen, bei denen Wassermoleküle zum Aufspalten von Molekülen verwendet werden.

Die größte Gruppe setzt sich aus 18 Subsubklassen zusammen, die aus den Hauptklassen der Transferasen und Hydrolasen stammen.

Die zweitgrößte Gruppe ist mit 17 Subsubklassen fast so groß wie die größte Gruppe und setzt sich ebenfalls aus Transferase- und Hydrolasereaktionen zusammen. Bei den Reaktionen dieser Gruppe wird auf der Eduktseite eine C-N und O-H Bindung gespalten. Eine C-O und N-H Bindung entsteht hingegen auf der Produktseite.

Die 14 Subsubklassen der drittgrößten Gruppe stammen aus den Hauptklassen der Transferasen, Hydrolasen und Lyasen. Damit ist sie in Bezug auf die Hauptklassenzusammensetzung am

heterogensten. Auf der Eduktseite dieser Reaktionen wird eine P-O und O-H Bindung gespalten. Auf der Produktseite entstehen wiederum eine P-O und O-H Bindung. Dieses Austauschmuster ist typisch für die Aufspaltung von Phosphatgruppen.
Die Subsubklassen der Oxidoreduktasen bilden keine Gruppen mit Subsubklassen anderer Hauptklassen. Sie bilden nur Gruppen untereinander. Häufig beinhalten die Gruppen der Oxidoreduktasereaktionen mehr Veränderungen als die übrigen Gruppen. Die Ursache hierfür sind häufig Kofaktoren, die bei vielen Oxidoreduktasereaktionen beteiligt sind. Diese führen häufig zu größeren Matrixdimensionen. Auch die Gruppe mit den zahlreichsten Veränderungen innerhalb der R-Matrix gehört zu den Oxidoreduktasen. Hierbei handelt es sich um eine Zweiergruppe aus den Subsubklassen 1.4.1.- und 1.5.1.-. Bei den Reaktionen dieser Subsubklassen sind NAD oder NADP als Kofaktor beteiligt (vgl. 3.4, Abbildung 64). Die meisten Subsubklassen der Oxidoreduktasen bilden keine Gruppen, sondern besitzen ein spezifisches Elektronentransfermuster.
Eine besondere Eigenschaft weisen die Subsubklassen der Gruppe 6 auf. Diese Gruppe setzt sich aus den 4 Subsubklassen der Subklasse 5.1.-.- zusammen. Bei den Reaktionen dieser Gruppe kommt es zu einer Veränderung in der absoluten Konfiguration eines Atoms von S nach R.

4.2.6 Bewertung des Programms und der Ergebnisse

Das im Rahmen dieser Arbeit entwickelte Programm erlaubt die automatische Berechnung von Reaktionsmatrizen (R-Matrizen), die das Elektronentransfermuster einer Reaktion beschreiben. Die Berechnung erfolgt lediglich auf der Grundlage der Edukt- und Produktmoleküle, welche dem Programm in Form von Textdateien übergeben werden. Nur anhand dieser Information werden die Strukturen der Moleküle verglichen, die Reaktionsstellen ermittelt und die R-Matrizen errechnet. Um einen Vergleich der R-Matrizen zu ermöglichen, werden die R-Matrizen kanonisiert und in R-Strings übersetzt. Obwohl einige Programme zur Modulation von Reaktionen entwickelt wurden, ist das vorliegende Programm zur R-Matrix-Berechnung auf Basis des Dugundji-Ugi-Modells bislang einzigartig. Der Grund hierfür ist, dass zunächst eine Atomzuordnung der Eduktatome auf die Produktatome generiert werden muss. Diese Hürde erwies sich als sehr viel komplexer als zunächst vermutet. Biochemische Reaktionen weisen eine große Vielfalt der unterschiedlichsten Reaktionen auf. Sie können viele verschiedene Edukt- und Produktmoleküle beinhalten mit komplexen Veränderungen in der Molekülstruktur. Ringsysteme können aufgespalten werden und einige Enzyme katalysieren mehrere aufeinander folgende Reaktionen. Häufig sind die Elektronentransfermuster unterbrochen und die R-Matrizen können so leicht größere Dimensionen annehmen.

Kern des Programms ist ein Algorithmus, der maximale gemeinsame Substrukturen erkennt. Dieser ist schnell und flexibel genug, um auch große Moleküle vergleichen zu können und auch komplexe Veränderungen in der Molekülstruktur erkennen kann. Mit Hilfe eines Systems, das maximale gemeinsame Substrukturen kombiniert, und eines Rankingsystems werden alle möglichen Atomzuordnungen errechnet. An dieser Stelle können aber gravierende Laufzeitprobleme entstehen, wenn es sehr viele Möglichkeiten gibt, die maximalen gemeinsamen Substrukturen zu kombinieren. Daher war es unumgänglich, auch heuristische Aspekte in das Verfahren zu integrieren.
Um einen Vergleich der R-Matrizen zu ermöglichen, musste ein neues Kanonisierungsverfahren entwickelt werden. Vorhergehende Algorithmen erwiesen sich für die Kanonisierung von R-Matrizen biochemischer Reaktionen als unzureichend. Die zugehörigen R-Graphen können fragmentiert und verzweigt sein. Wie das Atomzuordnungsproblem erwies sich auch das Kanonisierungsverfahren als sehr komplex.
In der Regel erfolgt die Berechnung der R-Matrizen und die Kanonisierung in weniger als einer Sekunde (Pentium-Rechner, 2,4 GHz). Sind größere Moleküle an der Reaktion beteiligt, kann die Rechenzeit auch im Sekunden- oder Minutenbereich liegen. Diese etwas längeren Laufzeiten stellen aber kein größeres Problem dar, in Hinblick darauf, dass die R-Matrix eines Enzyms nur einmal berechnet werden muss. Ebenfalls stellt die Genauigkeit, mit der die Atome zugeordnet werden, eine erhebliche Leistung dar, gemessen an der komplexen Ausgangsituation und der Vielfalt enzymatischer Reaktionen.

Mit Hilfe des Programms konnten die R-Matrizen für ein größeres Set von enzymatischen Reaktionen errechnet werden. Bereits die R-Matrixdimensionen der verschiedenen Subsubklassen lassen interessante Rückschlüsse auf die Reaktionen zu. Die R-Matrixdimension zeigt an, wieviele Atome an der Reaktion beteiligt sind. Beispielsweise ist eine R-Matrixdimension von 1 ein Anzeichen dafür, dass eine Veränderung in der absoluten Konfiguration eines Atoms erfolgt. Am häufigsten ist die R-Matrixdimension 4 vertreten. Bei vielen Reaktionen mit dieser Dimension werden 2 Bindungen auf der Eduktseite gespalten, während auf der Produktseite 2 Bindungen entstehen. Fast alle Subsubklassen der Transferasen, Hydrolasen und Lyasen besitzen diese Matrixdimension. Große R-Matrixdimensionen kommen überwiegend bei Oxidoreduktasereaktionen vor, bei denen Kofaktoren die Matrixdimension erhöhen. Eine ungerade Matrixdimension kann ein Anzeichen dafür sein, dass ein Atom an der Entstehung oder Spaltung von mehreren Bindungen beteiligt ist oder eine Änderung seiner Bindungsanzahl durch ein freies Elektronenpaar ausgleicht.
Die Subsubklassen erwiesen sich in der Regel als homogen in Bezug auf die Elektronentransfermuster ihrer Reaktionen. Dies war zu erwarten, da die EC-Klassifizierung

ausgehend von den Hauptklassen über die Subklassen bis hin zu den Subsubklassen immer spezifischer wird. Auf der Ebene der Subsubklassen sind die Reaktionen bereits so weit spezifiziert, dass sie weitgehend identisch sind. Die durch das Programm errechneten R-Matrizen zeigen deutlich, bei welchen Subsubklassen eine hohe Homogenität vorliegt und wo erhebliche Unterschiede in den Elektronentransfermustern einer Subsubklasse auftreten. Bei den Subsubklassen des Typs a.b.98.- oder a.b.99.- war eine geringe Homogenität zu erwarten. Allerdings zeigen die Auswertungen auch Unterschiede bei einigen Subsubklassen, für die man sonst eine höhere Homogenität erwarten würde. Eine andere interessante Fragestellung war, ob sich die Elektronentransfermuster der Reaktionen mit den Gruppen des EC-Klassifikationssystems decken. Zu diesem Zweck wurden die R-Matrizen der verschiedenen Subsubklassen verglichen und Subsubklassen mit identischem Elektronen-austauschmuster gruppiert. Es zeigte sich, dass 121 Subsubklassen bereits ein spezifisches Elektronentransfermuster besitzen. Die restlichen 107 Subsubklassen bilden hingegen Gruppen verschiedener Größe. Erstaunlicherweise beinhalten dabei einige Gruppen Subsubklassen aus verschiedenen Hauptklassen. So setzt sich die drittgrößte Gruppe aus Subsubklassen der Transferasen, Hydrolasen und Lyasen zusammen. Aus Perspektive des Dugundji-Ugi-Modells macht diese Zuordnung Sinn, da der Reaktionskern der zugehörigen Enzyme identisch ist. Bei allen Reaktionen werden Phosphatgruppen aufgespalten. Das von Dugundji und Ugi entwickelte System ist ein sehr rationales Modell, das sehr nüchtern und unvoreingenommen die wesentlichsten Eigenschaften einer Reaktion beschreibt. Die verschiedenen Gruppen mit identischen Elektronentransfermustern repräsentieren daher eine ganz neue Einteilung enzymatischer Reaktionen.

4.2.7 Ausblick

Im Zentrum der vorliegenden Arbeit stand die Entwicklung eines Programms zur automatischen Berechnung von Reaktionsmatrizen biochemischer Reaktionen. Die R-Matrizen repräsentieren aber nicht nur den Kern einer Reaktion. Sie enthalten bereits die wichtigste Information, die benötigt wird, wenn Reaktionen automatisch beschrieben werden sollen. Der schwierigste Schritt bei der automatischen Charakterisierung von Reaktionen ist die Identifizierung der Reaktionsstellen. Auf Grundlage dieser Information ist es einfach, die Reaktionen sehr viel präziser zu charakterisieren. Dies kann zum Beispiel hilfreich sein, wenn das System für eine Vorhersage der Subsubklasse eingesetzt werden soll.

Wie im Ergebnisteil beschrieben, besitzen 121 Subsubklassen bereits ein spezifisches Elektronentransfermuster. Bei neuen Vertretern dieser Subsubklassen ist damit schon eine

Subsubklassenvorhersage anhand ihres Elektronentransfermusters möglich. Anders ist dies bei den restlichen 107 Subsubklassen, die nicht über ein spezifisches Elektronentransfermuster verfügen. Um hier eine Subsubklassenvorhersage zu ermöglichen, müssen weitere Informationen hinzugezogen werden. Dies können Informationen über die Umgebungen der Atome sein, die an der Reaktion beteiligt sind. Wenn bekannt ist, welche funktionellen Gruppen durch die Reaktion ineinander überführt werden, lässt sich die Reaktion besser charakterisieren und die Subsubklassenvorhersage verbessern.
Neben einer Subsubklassenvorhersage könnte das Dugundji-Ugi-Modell dazu verwendet werden, das bestehende EC-Klassifikationsystem zu verbessern. Die Elektronentransfermuster der Enzyme könnten dazu beitragen, die Subsubklassen feiner zu untergliedern oder Subsubklassen mit gleichen Elektronentransfermustern zu größeren Gruppen zusammen zu fassen.

Um die R-Matrix-Berechnung zu ermöglichen, war die Entwicklung von Algorithmen erforderlich, die ebenso zum Vergleich von Metaboliten oder zur Analyse von metabolischen Pfaden verwendet werden können. Mit Hilfe des *c*-MCS-Algorithmus ist es möglich, die Substratmoleküle verschiedener Enzyme zu vergleichen. Verwenden zwei Enzyme ähnliche Substratmoleküle, kann dies ein Hinweis darauf sein, dass sie eine ähnliche Reaktion katalysieren.
Der *c*-MCS-Algorithmus wird in Zusammenwirken mit dem *c*-MCS-Kombinationssystem und dem Rankingsystem dazu verwendet, eine atomare Zuordnung für biochemische Reaktionen zu generieren. Hierdurch wird eine genaue Zuordnung der Strukturen der Eduktmoleküle auf die Strukturen der Produktmoleküle erzeugt. So kann der Weg eines einzelnen Atoms über mehrere Reaktionen hinweg nachvollzogen werden. Bislang stellt es ein erhebliches Problem dar, den Weg eines Moleküls durch den Stoffwechsel zu verfolgen. Im Grunde wird jeder Metabolit umgewandelt oder aufgespalten. Molekülstrukturen werden auf andere Moleküle übertragen und folgen verschiedenen Verzweigungen innerhalb des Netzwerkes von metabolischen Pfaden. Bisherige Arbeiten (Rahman, 2005) verwenden ungenaue Verfahren, wie Fingerprint Methoden im Zusammenwirken mit dem Tanimoto-Ähnlichkeitskoeffizienten, um Molekülstrukturen im Stoffwechsel zu verfolgen.
Das im Rahmen dieser Arbeit entwickelte Atomzuordnungsverfahren kann daher zu erheblichen Verbesserungen bei der Analyse von metabolischen Pfaden führen.

5 Literatur

Bayada, D. M., Simpson, R. W., Johnson, J. P., (1992) An Algorithm for the Multiple Common Subgraph Problem, *J. Chem. Inf. Comput. Sci.*, **32**, 680 – 685

Bauer, J., (1989) IGOR2: a PC-Program for Generating New Reactions and Molecular Structures, *Tetrahedron Comput. Methodol*. **2**, 269-280.

Brandt J., Bauer J., Frank R. M., Scholley A., (1981) Classification of Reactions by Electron Shift Patterns, Chemica Scripta. **18**, 53-60.

Brandt J., von Scholley A., (1983) An efficient algorithm for the computation of the canonical numbering of reaction matrices. *Computers & Chemistry* **7(2)**, 51-59.

Brint A. T., Willet P., (1987) Algorithms for the identification of three-dimensional maximal common substructures , *J. Chem. Inform. Comput. Sci.* **2** 311-320

Bron C. and Kerbosch J., (1973) Algorithm 457 - Finding all cliques of an undirected graph. *Communications of the ACM* **16**, 575-577.

Cahn R.S., C.K. Ingold C.K., Prelog V., (1966) Spezifikation der molekularen Chiralität, *Angewandte Chemie*, **78**, 413–447.

Dugundji J., and Ugi I., (1973) An algebraic model of constitutional chemistry as a basis for chemical. computer programs, *Topics Current Chem.* **39**, 19–64.

Durand P., Pasari R., Baker J., Tsai C. (1999) An Efficient Algorithm for Similarity Analysis of Molecules, *Internet Journal of Chemistry*, 1999 - ijc.com Article 17.

Eisenmesser EZ, Bosco DA, Akke M, Kern D., (2002) Enzyme dynamics during catalysis. *Science*, 22;295(5559):1520–3.

Enzyme Nomenclature
Classification and Nomenclature of Enzymes by the Reactions they Catalyse, 1992 http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html

Fersht, A (1985) Enzyme Structure and Mechanism (2nd ed) p50–52 W H Freeman & co, New York

Fontain, E., Reitsam, K., (1991) The Generation of Reaction Networks with RAIN. 1. The Reaction Generator, *J. Chem. Inf. Comput. Sci.* **31**, 96-101.

García G. C., Ruiz I. L., Gómez-Nieto M. Á., (2004) Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction Based Algorithm, *J. Chem. Inf. Comput. Sci.* **44** (1), 30 -41.

Garcia-Viloca M., Gao J., Karplus M., Truhlar D. G. (2004). How enzymes work: analysis by modern rate theory and computer simulations, *Science* 303 (5655): 186–195.

Gerhards L., Lindenberg W., (1979) Clique detection for non-directed graphs: two new algorithms, Computing 21, 295-322.

Hart H., Craine L. E., Hart D. J., (2002) Organische Chemie, 2. Auflage, Wiley VCH Verlag GmbH

Hartnell B., Rall D., (1998) Domination in Cartesian Products: Vizing's Conjecture. In Domination in Graphs--Advanced Topics (Ed. T. W. Haynes, S. T. Hedetniemi, and P. J. Slater). New York: Dekker, pp. 163-189, 1998.

Hattori M., Okuno Y., Goto S. and Kanehisa M., (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *Journal of the American Chemical Society* **125**, 11853-11865.

Historical Introduction - Report from the introduction to Enzyme Nomenclature, 1992 http://www.chem.qmul.ac.uk/iubmb/enzyme/history.html

Johnston H. J., (1976) Cliques of a graph – variations of the Bron-Kerbosch algorithms, *Int. J. Comput. Inform Sci*. , **5**, 209 – 238.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y., (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**(Database issue):D480-4.

Keith F. Tipton, Sinéad Boyce (2000) History of the enzyme nomenclature system. *Bioinformatics* **16** (1), 34-40.

Koch I., (2001) Enumerating all connected maximal common subgraphs in two graphs. *Theor. Comput. Sci.* **250**(1-2): 1-30.

Koshland D. E., (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis, *Proc. Natl. Acad. Sci.* **44** (2): 98–104.

Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M., (2004) Computational assignment of the EC-numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society* **126**, 16487-16498.

Levi G., (1972) A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *Calcolo* **9**, 341-352.

Li E, Hristova K., (2006) Role of receptor tyrosine kinase transmembrane domains in cell signaling and human pathologies, *Biochemistry* **45** (20): 6241-51.

Lüdke, S., (2002) Kodierung von Enzymreaktionen, Diplomarbeit am Institut für Biochemie, Universität zu Köln

Marialke J., Korner R., Tietze S., Apostolakis J., (2007) Graph-Based Molecular Alignment (GMA). *Journal of Chemical Information and Modelling* **47**(2):591-601.

McGregor J. J., (1982), Backtrack Search Algorithms and the Maximal Common Subgraph Problem. Softw., Pract. Exper. **12**(1): 23-34.

Morgan David O., (2007) The Cell Cycle: Principles of Control. New Science Press: London.

Oh M., Yamada T., Hattori M., Goto S., Kanehisa M., (2007) Systematic Analysis of Enzyme-Catalyzed Reaction Patterns and Prediction of Microbial Biodegradation Pathways. *J. Chem. Inf. Mode.*, Web Release Date: May 22, 2007

Prelog V., Helmchen G., (1982) Grundlagen des CIP-Systems und Vorschläge für eine Revision, *Angewandte Chemie*, **94**, 614-631

Rahman S.A., Advani P., Schunk R., Schrader R., Schomburg D. (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC), *Bioinformatics*, **21** (7) 1189-1193

Raymond J. W., Willett P., (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design* 16 (Number 7): 521-533

Rodbell M., (1980) The role of hormone receptors and GTP-regulatory proteins in membrane transduction, *Nature* **284** (5751): 17-22.

Schmidt, S., Sunyaev, S., Bork, P. & Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? *Trends Biochem Sci* **28**, 336-41.

Schomburg I., Chang A., Ebeling C., Gremse M., Heldt C., Huhn G., Schomburg D., (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*, **32**(Database issue):D431-3.

Srivastava D. K., Bernhard S. A., (1986). Metabolite transfer via enzyme-enzyme complexes. *Science* **234**, 1081-1086

Ugi I., Bauer J., Blomberger C., Brandt J., Dietz A., Fontaine E., Gruber B., Scholley-Pfab A., Senff A., Stein N., (1994) Models, concepts, theories, and formal languages in chemistry and their use as a basis for computer assictance in chemistry, J. chem. inf. comput. sci. **34**, 3-16.

Donald Voet, Judith G. Voet (2004): Biochemistry, 3. Auflage. John Wiley & Sons Inc., London

WEBB EC-(1990) Enzyme Nomenclature – Recommendations -1984 - Supplement-3 - Corrections and Additions *European Journal of Biochemistry* **187** (2), 263-281.

Anhang:

Auf der CD befinden sich die Programme „cMCS Searcher“ und „Reaction Matrix Calculator“:

cMCS Searcher:

Berechnet die zusammenhängenden maximalen gemeinsamen Substrukturen (cMCS) von 2 Molekülen. Der Algorithmus ist eine Kombination aus einer Variante des Bron-Kerbosch-Algorithmus und dem McGregor-Algorithmus.
Die cMCS werden in Form von Atomzuordnungen gespeichert und in einer Liste von Integer Vektoren abgelegt (Liste final_MAPPINGS der Klasse MCS_Search). Die Ausgabe der Atomzuordnungen erfolgt mitunter als Standardausgabe in die Konsole.

Reaction Matrix Calculator:

Das Programm berechnet die Reaktionsmatrizen von chemischen oder biochemischen Reaktionen. Die Berechnung erfolgt auf Basis der Edukt- und Produktmoleküle, die dem Programm als Parameter übergeben werden. Die Moleküle müssen in Form des MDL Formates als Molfiles vorliegen. Nacheinander werden die Edukt- und Produktmolekuele paarweise durch den cMCS-Algorithmus verglichen, eine Atomzuordnung für die Reaktion generiert, die Reaktionsmatrizen errechnet, die Reaktionsmatrizen kanonisiert und in Reaktionsstrings umgewandelt.
Die Reaktionsmatrizen werden in das Textfile „final_solutions.txt“ und die Reaktionsstrings in das Textfile „Canonical_Vectors.txt“ ausgegeben.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Dietmar Schomburg betreut worden.

Teilpublikationen: keine
Köln, den 11.02.2008

Lebenslauf:

Name:	Leber	
Vorname:	Markus	
geboren:	07.08.1972 in Winterberg	
Adresse:	Heisterbacherstraße 20 53332 Bornheim	
Telefon:	02222/82656	
Schule:	01.08.1979	Einschulung an der Kath. Grundschule 59955 Winterberg (Züschen)
	07.01.1980	Gemeinschafts-Grundschule Hersel Rheinstraße 166 - 53332 Bornheim
	August 1983	Collegium Josephinum Bonn Gymnasium in Trägerschaft des Provinzialates der Redemptoristen e.V. Köln
	August 1989	Ernst Moritz Arndt - Gymnasium
	Juni 1992	Zeugnis der allgemeinen Hochschulreife
Zivildienst:	Juli 1992 - September 1993	Wohnstift Augustinum Bonn (Altenpflege)
Studium:	Oktober 1993	Studium der Biologie Rheinische Friedrich-Whilhelms-Universität Bonn
	1999	Diplomarbeit Kekulè Institut für Organische Chemie und Biochemie der Universität Bonn
	27.12.1999	Diplom in Biologie
	01.2000 - 03.2000	Beschäftigung am Kekulè Institut Bonn
	10.2000 - 02.2001	Praktikum „Molekulare Bioinformatik“ in Bonn
	02.2001 - 12.2001	Bioinformatikkurs Akademie für Weiterbildung in Heidelberg
	02.2002 - 03.2003	Bioinformatikstudiengang Cologne University BioInformatics Center (CUBIC)
Promotion:	Seit Juni 2003	Promotion Cologne University BioInformatics Center (CUBIC)