Erweiterte Identifizierung, automatische Generierung und Analyse von konservierten Sequenzmustern und vergleichende Analyse enzymatischer Reaktionen unter Verwendung von homologen Enzymdomänen

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Adrian Welfle

aus Düsseldorf

Köln, 2008

Berichterstatter: Prof. Dr. D. Schomburg

Prof. Dr. R. Schrader

Tag der mündlichen Prüfung: 26.06.2008

Danksagung:

Ich danke Herrn Prof. Dr. D. Schomburg für die Bereitstellung des Themas und die ständige Möglichkeit der fachlichen Diskussion. Ein herzliches Dankeschön auch an Prof. Dr. R. Schrader für die Übernahme des Zweitgutachtens.

Ich danke zudem Markus Leber, der zum Gelingen der Arbeit beitrug.

Eine Danksagung kann nie vollständig sein. Ich danke daher allen, die ihren Namen in dieser Danksagung nicht erwähnt sehen.

Schließlich möchte ich mich recht herzlich bei meiner Familie, insbesondere bei meiner Mutter, bedanken, ohne sie diese Arbeit nicht möglich gewesen wäre.

„Wir geben niemals auf, öffnen neue Türen und entwickeln neue Ideen, weil wir neugierig sind. Die Neugier ist es, die uns neue Wege beschreiten lässt. Gib niemals auf!"

Walt Disney

Inhaltsverzeichnis

Abkürzungsverzeichnis

3D	dreidimensional
A	
Abb.	Abbildung
ADP	Adenosindiphosphat
ATP	Adenosintriphosphat
AS	Aminosäure
B	
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrix
bzw.	Beziehungsweise
C	
c-MCS	connected-Most Common Subgraph
CPU	Central Processor Unit
D	
Da	Dalton
d.h.	Das heißt
DNA	Desoxyribonucleinsäure
DVD	Digital Versatile Disc
E	
EC	Enzyme Commission
Evalue	Expectation Value
E-Wert	Deutsche Bezeichnung für Expectation Value
F	
F-P	Falsch-Positiv
G	
GDP	Guanosindiphosphat
GTP	Guanosintriphosphat
Ggf.	Gegebenenfalls
H	
H-Bindung	Wasserstoffbrückenbindung
K	
kDa	Kilodalton
M	
MCS	Most Common Subgraph
MB	Megabyte
MHz	Megahertz

N

NADH	Nicotinsäureamidadenindinucleotid (reduziert)
NADPH	Nicotinsäureamidadenindinucleotidphosphat (reduziert)
NCBI	National Center of Biotechnology Information
NMR	Nuclear Magnetic Resonance

P

PDB	Protein Data Bank
PGH	Phosphoglycolohydroxamat

R

RAM	Random Access Memory
R-Matrix	Reaktionsmatrix
R-P	Richtig-Positiv

S

Sog.	Sogenannt

T

Tab.	Tabelle

U

usw.	Und so weiter

V

vgl.	Vergleiche

Z

z.B.	Zum Beispiel

Abkürzungen der Aminosäuren

A	Ala	Alanin
B		Aspartat oder Asparagin
C	Cys	Cystein
D	Asp	Aspartat
E	Glu	Glutamat
F	Phe	Phenylalanin
G	Gly	Glycin
H	His	Histidin
I	Ile	Isoleucin
K	Lys	Lysin
L	Leu	Leucin
M	Met	Methionin
N	Asn	Asparagin
P	Pro	Prolin
Q	Gln	Glutamin
R	Arg	Arginin
S	Ser	Serin
T	Thr	Threonin
U		Selenocystein
V	Val	Valin
W	Trp	Tryptophan
X		Unbekannte oder Nichtstandard-Aminosäure
Y	Tyr	Tyrosin
Z		Glutamat oder Glutamin

Abstract

Enzymes are biomolecules that catalyze chemical reactions in living organisms. Almost all processes in a biological cell need enzymes in order to occur at significant rates and almost all enzymes are proteins. Although enzymes are able to catalyze different reactions, they may contain similar modular domains conserved throughout evolution. Domains are the structural, functional and evolutionary units of proteins.

The International Union of Biochemistry and Molecular Biology (IUBMB) classifies enzymes into six groups. The classification of enzymes is based on their catalyzed reactions and not on similar domains or sequences. As the number of protein sequences in public databases grows rapidly with the progress of experimental technologies in molecular biology, the need for accurate protein annotation from amino acid sequences only is a central problem in computational biology.

In this work, the cluster analysis as a widely used method in computational biology was used to group sequences into meaningful clusters according to their sequence similarities. The result of this analysis and the construction of sequence patterns should help in the understanding of the relationship between sequence similarity and similar function.

First, all sequences of currently available sequences that contain at least one complete EC-Number were collected. The result of all-vs-all BLAST alignments was used, to assign the domain structure of the analyzed sequences. Depending on the E-value of these alignments, domains that share sequence similarity were classified into groups of homologous proteins. From certain clusters, sequences were taken to construct sequence patterns. The quality of these patterns was tested by searching for True-Positive or False-Positive hits. A hit was defined as True-Positive, if the hit contains the same EC-Number as the pattern. The resulting patterns were also compared to patterns derived from the PROSITE database.

Additionally, an algorithm, which determines maximal common substructures of molecules, was used to compare molecules, which were involved during catalytic reactions by the compared enzymes. Finally, the result of the cluster analysis based on sequence similarity was compared with the result of a cluster analysis based on enzymes, which were grouped because of identical reaction matrices.

118947 sequence patterns were constructed and their fitness was tested. Most of these patterns were constructed from up to ten sequences at high E-values. Examples showed, that generally

amino acids, which are responsible for the catalytic activity of enzymes or those which are important in assuring a right 3D conformation, are highly conserved.

The comparison of molecules, which are involved during catalysis of clustered enzymes showed, that most enzymes use identical or very similar molecules. Depending on the E-value, the occurrence of identical molecules being used during catalysis decreases with ascending E-value. Additionally, the cluster comparison based on sequence similarity with other clusters based on identical reaction matrices, showed, that the number of sequences, which were grouped by both methods, decrease with ascending E-value.

Zusammenfassung

Enzyme sind Biomoleküle, die chemische Reaktionen in lebenden Organismen katalysieren. Nahezu alle Reaktionen in einer lebenden Zelle benötigen Enzyme, damit chemische Reaktionen in angemessener Zeit ablaufen. Annähernd alle Enzyme sind Proteine. Obwohl Enzyme in der Lage sind, unterschiedliche Reaktionen zu katalysieren, können sie gleiche Domänen enthalten, die sich während der Evolution konserviert haben. Domänen sind die strukturellen, funktionellen und evolutionären Einheiten von Proteinen.

The International Union of Biochemistry and Molecular Biology teilt Enzyme in sechs Klassen ein. Die Einteilung wird anhand der katalysierten Reaktion vorgenommen, nicht anhand gleicher Domänen oder Sequenzen. Da die Anzahl sequenzierter Proteine aufgrund von innovativen Sequenzierungstechnologien schnell wächst, ist die korrekte Annotation von Enzymen anhand reiner Sequenzinformation ein zentrales Problem in der Bioinformatik.

In dieser Arbeit wurde die Clusteranalyse als etablierte und häufig genutzte Methode in der Bioinformatik dazu genutzt, Sequenzen anhand ihrer Sequenzähnlichkeit zu bedeutsamen Clustern zu gruppieren. Das Ergebnis dieser Analyse und die Erstellung von Sequenzmustern sollen helfen, die Frage zu beantworten, inwiefern es möglich ist, von Sequenzähnlichkeit auf gleiche Funktion zu schließen.

Zunächst wurden alle derzeit verfügbaren Enzymsequenzen, die mindestens eine vollständige EC-Nummer tragen, gesammelt. Das Ergebnis von all-vs-all Alignments wurde dazu genutzt, die Domänenstruktur der analysierten Enzyme zu bestimmen. Abhängig vom E-Wert dieser Alignments, wurden Cluster aus homologen Domänen gebildet. Aus bestimmten Clustern wurden Sequenzen entnommen, um daraus Sequenzmuster zu erstellen. Die Qualität dieser Muster wurde durch Suche nach Richtig-Positiven und Falsch-Positiven Treffern getestet. Ein Treffer wird als Richtig-Positiv definiert, wenn der Treffer die gleiche EC-Nummer enthält, wie das Muster. Die erstellten Muster wurden mit Mustern der PROSITE-Datenbank verglichen.

Zusätzlich wurde ein Algorithmus, der die größte gemeinsame Teilstruktur bestimmt, dazu genutzt, um Moleküle, die bei geclusterten Enzymen bei der Katalyse beteiligt sind, miteinander zu vergleichen. Reaktionsmatrizen wurden auf diese Weise erstellt. Schließlich

wurde das Ergebnis der Clusteranalyse, die aufgrund Sequenzähnlichkeit basiert, mit dem Ergebnis der Clusteranalyse verglichen, die aufgrund identischer Reaktionsmatrizen basiert.

118947 Sequenzmuster wurden erstellt und deren Qualitäten bestimmt. Der größte Teil der Muster wurde aus bis zu zehn Sequenzen bei hohen E-Werten erstellt. Beispiele zeigten, dass Aminosäuren, die für die katalytische Aktivität oder für die Gewährleistung der korrekten 3D Konformation verantwortlich sind, hochkonserviert sind.

Der Vergleich der Moleküle, die bei geclusterten Enzymen beteiligt sind, zeigte, dass die meisten Enzyme identische oder sehr ähnliche Moleküle nutzen. Abhängig vom E-Wert, nimmt die Anzahl von identischen Molekülen bei verglichenen Reaktionen mit ansteigendem E-Wert ab. Zusätzlich konnte bei dem Vergleich des Ergebnisses der Clusteranalyse, die auf Sequenzähnlichkeit basiert, mit dem Ergebnis der Clusteranalyse, die auf gleichen Reaktionsmatrizen basiert, gezeigt werden, dass die Anzahl der Enzyme, die in beiden Clusteranalysen gruppiert wurden, mit steigendem E-Wert abnimmt.

1 Einleitung

1.1 Proteine und Enzyme

Proteine (griechisch protos = “erstes, wichtiges“) sind biologische Makromoleküle mit einer immensen Vielfalt in Form und Funktion. Haare und Muskelfibrillen, Kollagen, Myoglobin und Seidenfibroin sind Proteine unterschiedlichen Aufbaus. Aber auch Hormone, Pilzgifte, Antikörper und zahllose weitere wichtige Substanzen bestehen aus Proteinen [125].

Die zahlenmäßig größte Proteingruppe bilden die Enzyme. Ohne Enzyme ist Leben auf der Erde nicht vorstellbar, denn Enzyme beschleunigen verschiedenste Reaktionen, die unter physiologischen Bedingungen sonst in einem lebenswidrigen Zeitrahmen ablaufen würden. Enzyme, die Katalysatoren bei anabolen und katabolen Reaktionen in lebenden Organismen, beschleunigen Reaktionen um mindestens den Faktor 10^5, können aber auch die Reaktionsgeschwindigkeit um 14 Größenordnungen (z.B. bei Urease) durch Senkung der Aktivierungsenergie beschleunigen [125]. Die Leistung der Enzyme lässt sich veranschaulichen, wenn man sich vor Augen führt, dass der Mensch mit Hilfe mehrerer Enzyme das äußerst stabile Molekül Glucose innerhalb kürzester Zeit metabolisiert und lebenswichtige Energie gewinnt. Dabei sind Enzyme nötig, die in gut organisierter Abfolge in die Glucoseverwertung eingreifen.

Enzyme haben ein ph- und Temperaturoptimum; sie arbeiten aber auch bei Bedingungen, die weit außerhalb der Optima liegen. Diese spezialisierten biologischen Katalysatoren sind oft spezifischer und leistungsfähiger, als viele nicht-biologische Katalysatoren. Aufgrund der katalytischen Leistung finden Enzyme Anwendung in der Landwirtschaft, Lebensmittelverarbeitung, chemischen Industrie und Medizin [122]. Im Laufe der stetig fortschreitenden Struktur- und Funktionsaufklärung, erlangt auch die biotechnologische Modifikation der Enzyme in biologischen Reaktionen in der Industrie, zum Beispiel bei der Fermentation, zunehmende Bedeutung. Ziel ist häufig, die Enzymsequenz zu modifizieren, um das Temperaturoptimum des Enzyms zu erhöhen [123].

Trotz einer erstaunlichen Komplexität nutzen Enzyme wenige Reaktionsprinzipien. Sie bestehen, bis auf wenige Ausnahmen, aus einem Satz von 20 verschiedenen Bausteinen, den Aminosäuren (eine Übersicht der Aminosäuren findet sich zu Beginn dieser Arbeit). Die Primärstruktur eines Proteins ist die Anordnung der Aminosäuren zu einer Sequenz. Diese faltet sich und kann typische Sekundärstrukturen, z.B. das Beta-Faltblatt oder die Alpha-Helix bilden. Weiterführende Strukturen im Raum sind die Tertiärstruktur (dreidimensionale

Anordnung der Polypeptidkette) bzw. die Quartärstruktur (räumliche Struktur von Polypeptidketten zu funktionellen Untereinheiten) [126].

1.2 Die Proteindomäne

Neben den vier eben genannten Organisationseinheiten wurde eine Unterstruktur, die Proteindomäne, identifiziert. Sie wird von Teilen einer Polypeptidkette gebildet und kann sich unabhängig zu einer kompakten und stabilen Struktur falten. Sie ist die strukturelle, funktionelle und evolutionäre Einheit von Proteinen [69]. Eine Domäne enthält circa 40 bis 350 Aminosäuren und ist die Einheit, aus der viele größere Proteine aufgebaut sind. Dabei haben verschiedene Domänen eines Proteins oft unterschiedliche Funktionen. Große Proteine können aus mehreren Dutzend Domänen bestehen, kleinste Proteine enthalten nur eine Domäne [127].

Abbildung 1-1: Domänen eines Enzyms: Serin Hydroxymethyltransferase aus *Bacillus stearothermophilus*. PDB Datei 1KKP. C-terminale Domäne orange. Die große N-terminale Domäne kann noch in Subdomänen unterteilt werden: N-terminale Subdomäne blau, große Pyridoxalphosphat-bindende Subdomäne grün.

Eine exakte Charakterisierung des Begriffs der Proteindomäne ist schwierig, da diese Bezeichnung in der Fachliteratur unterschiedlich interpretiert wird. Grundlage für die Definition ist teilweise die Struktur der Domäne; andere definieren die Domäne als typische

Unterstruktur, die in verschiedenen Proteinen vorkommt [70]. Auch die Tatsache, dass verschiedene Domänen in Subdomänen eingeteilt werden (vgl. Abbildung 1-1), verdeutlicht die Schwierigkeit der Abgrenzung des Domänenbegriffs.

Die individuelle Anordnung der Aminosäuren innerhalb einer Aminosäurekette ist für die komplexe dreidimensionale Struktur von Enzymen (und Proteinen allgemein) und deren Funktion verantwortlich. Eine Untersuchung der Primärsequenz ist für die Aufklärung der Funktionsweise wichtig. Die Analyse der Anordnung der Aminosäurekette im Raum kann zur Aufklärung der Enzymfunktion im Detail beitragen.

1.3 Die Funktionsweise der Enzyme: eine kurze Einführung

Enzyme sind sehr leistungsfähig und spezifisch für ein bestimmtes Molekül (Substrat). Wie kommen diese Eigenschaften zustande? Da sich diese Dissertation mit der Analyse von Enzymen beschäftigt, gibt der folgende Abschnitt eine kurze Einführung in die Funktionsprinzipien von Enzymen.

Enzyme binden das Substrat in einer definierten Tasche des Moleküls, dem aktiven Zentrum. Speziell angeordnete Aminosäuren im aktiven Zentrum nehmen Kontakt zum Substratmolekül auf, der Enzym-Substrat-Komplex entsteht. Durch multiple Wechselwirkungen zwischen Enzym und Substrat wird freie Enthalpie freigesetzt, die sowohl für die Spezifität, als auch für die Leistungsfähigkeit der Enzyme verantwortlich ist. Durch die geschützte Lage des aktiven Zentrums verläuft die katalytische Reaktion weitgehend getrennt vom Lösungsmittel ab und wird dadurch energetisch begünstigt. Viele, wenn nicht gar alle Wasserstoffbrücken zwischen dem Substrat und Wasser werden getrennt. Dadurch wird die Aktivierungsenergie gesenkt und die Reaktion beschleunigt [125]. Abbildung 1-2 zeigt das Enzym Hexokinase, das bei der Metabolisierung von Glucose beteiligt ist. ATP und Glucose, die am Molekül gebunden sind, wurden farbig markiert.

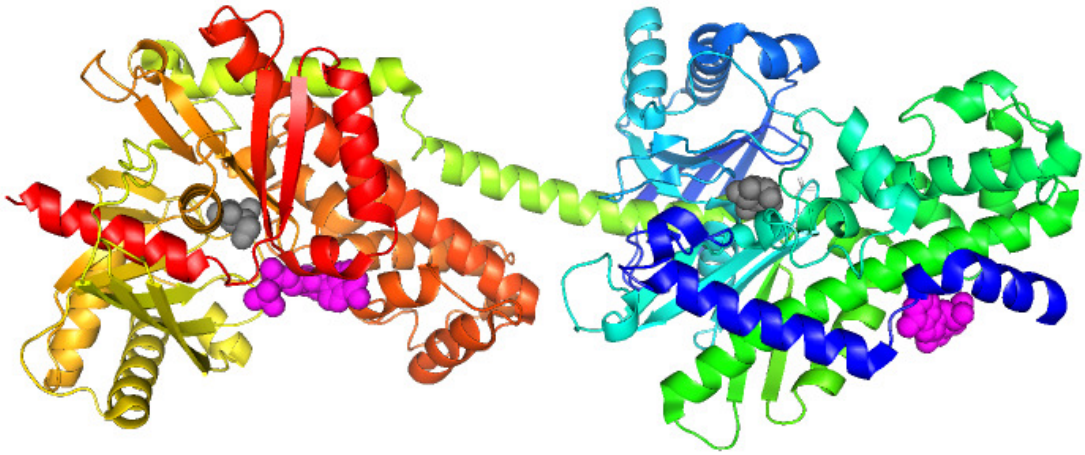


Abbildung 1-2: Hexokinase aus *Homo sapiens*, PDB 1DGK, Dimer. ATP magenta, Glucose grau.

Um eine optimale Bindung zwischen dem aktiven Zentrum und dem Substrat zu gewährleisten, muss das aktive Zentrum komplementär zum Substratmolekül sein. Man spricht in diesem Zusammenhang vom Schlüssel-Schloss-Prinzip, das erstmals 1946 von Linus Pauling erarbeitet wurde [125]. Dabei muss beachtet werden, dass im aktiven Zentrum der Übergangszustand des Substrats während der Reaktion stabilisiert wird, damit die Reaktion optimal beschleunigt werden kann. Die Bindungsenergie zwischen Enzym und Substrat im Übergangszustand leistet einen entscheidenden Beitrag zur Leistungsfähigkeit des Enzyms.

1.4 Die EC-Nomenklatur der Enzyme

Zu Beginn der wissenschaftlichen Analyse der Enzyme war es üblich, dass der Entdecker eines Enzyms für die individuelle Namensgebung verantwortlich war. Häufig beliebig oder durch das Anfügen der Endung „-ase“ an den entsprechenden Namen, wurden Enzyme ohne klare Regeln benannt. So katalysiert zum Beispiel die DNA-Polymerase die Synthese von DNA aus Nucleotiden. Viele Enzyme erhielten auf diese Weise nicht eindeutige Namen, so dass die Bezeichnungen der Enzyme nicht auf die katalysierten Reaktionen schließen ließen. Viele Enzyme trugen mehrere verschiedene Bezeichnungen, unterschiedliche Enzyme trugen identische Namen. Die willkürliche Benennung und die ständig wachsende Zahl neu entdeckter Proteine hatten zur Folge, dass die Enzymnomenklatur so unübersichtlich wurde, dass eine Einteilung der Enzyme in ein einheitliches System nötig war. 1961 schlug die *Enzyme Commission* (EC) der *International Union of Biochemistry* (IUB; heute *International Union of*

Biochemistry and Molecular Biology, IUBMB) internationale Richtlinien zur Benennung und Klassifizierung von Enzymen vor. Nach diesem System erhält ein Enzym eine systematische und hierarchische vierstellige EC-Nummer (Einteilung in Haupt- und Subklasse und weitere Differenzierung), sowie einen systematischen Namen. Vorhandene Namen werden gesammelt und aufgeführt. Die Eingliederung von Enzymen in die EC-Nomenklatur erfolgt nach dem Typ der katalysierten Reaktion. Der Stand der aktuellen EC-Nummern unterliegt ständiger Überarbeitung [128].

Anhand der folgenden Reaktion soll das EC-System erklärt werden:

ATP + D-Hexose ↔ ADP + D-Hexose-6-Phosphat

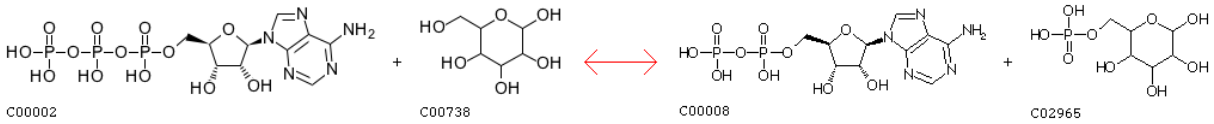


Abbildung 1-3: Katalysierte Reaktion des Enzyms ATP: D-Hexose-Phosphotransferase.

Der systematische Name des Enzyms, das diese Reaktion katalysiert, lautet:

ATP: D-Hexose-Phosphotransferase. Es trägt die EC-Nummer **2.7.1.1**, die sich wie folgt zusammensetzt:

(Die EC-Nummer wurde zur Verdeutlichung der Einteilung untereinander geschrieben)

- **2** → Die erste Ziffer teilt das Enzym in eine der sechs Enzymklasse ein. Hier: Gruppe der Transferasen.
- **7** → Die zweite Ziffer zeigt die Subklasse an. Hier: Transfer von Gruppen, die Phosphor enthalten.
- **1** → Subsubklasse: Das Enzym enthält eine Hydroxylgruppe als Akzeptor.
- **1** → Laufende Nummer: Glucose dient als Akzeptor der Phosphatgruppe.

Diese Einteilung gilt für mehrere tausend bekannte Enzyme. Eine vollständige Beschreibung aller EC-Klassen würde über den Rahmen dieser Arbeit hinausgehen. Die folgende Tabelle zeigt jedoch alle EC-Hauptklassen und den Typ der katalysierten Reaktion. Da sich die Grundlage dieser Arbeit und auch die ausgewählten Beispiele auf die EC-Nomenklatur berufen, erscheint diese Tabelle und die darauffolgende detailierte Beschreibung der EC-Klassen umso wichtiger.

EC-Klasse		**Typ der katalysierten Reaktion**
1	Oxidoreduktasen	Elektronentransfer (Hydrid-Ionen oder H-Atome)
2	Transferasen	Gruppentransfer-Reaktionen
3	Hydrolasen	Hydrolytische Spaltung
4	Lyasen	Abspaltung von Gruppen nach einem nicht-hydrolytischen Mechanismus
5	Isomerasen	Transfer von Gruppen innerhalb von Molekülen, dabei Bildung von Isomeren
6	Ligasen	Verknüpfung neuer Bindungen unter Spaltung energiereicher Phosphatbindungen

Tabelle 1-1: Die Einteilung der Enzyme in die sechs Hauptklassen gemäß EC-Nomenklatur.

EC-Klasse 1: Oxidoreduktasen

- Zu dieser Klasse gehören alle Enzyme, die Oxidoreduktionen katalysieren. Das Substrat, das oxidiert wird, dient als Wasserstoff- oder Elektronendonor. Bei der Namensgebung wird meist das Donormolekül, dann das Akzeptormolekül genannt, gefolgt von „Oxidoreduktase“. Ist ein Akzeptor O_2, wird das Enzym „Oxidase“ genannt.

EC-Klasse 2: Transferasen

- Transferasen sind Enzyme, die eine Atomgruppe, z.B. eine Methyl- oder Glycosylgruppe, von einem Donormolekül auf einen Akzeptor transferieren. Die Enzyme dieser Klasse werden meist nach dem Schema „Donor: Akzeptor Gruppentransferase“ oder „Donor (Akzeptor) Gruppentransferase“ benannt. In vielen Fällen ist der Donor ein Co-Faktor (Co-Enzym), der die zu transferierende Gruppe trägt.

EC-Klasse 3: Hydrolasen

- Die Enzyme dieser Klasse katalysieren die Spaltung verschiedener Bindungen unter Beteiligung von Wasser. Einige Enzyme sind sehr unspezifisch, so dass es in vielen Fällen nicht einfach ist zu unterscheiden, ob zwei untersuchte Enzyme, die anhand ihrer katalysierten Reaktion klassifiziert werden, identisch sind. Der systematische Name enthält immer „Hydrolase“. In Trivialnamen wird der Enzymname meist aus Substrat, gefolgt von dem Suffix „-ase“, aufgebaut.

EC-Klasse 4: Lyasen

- Lyasen spalten die Bindungen C-C, C-O, C-N und weitere Bindungen, ohne Beteiligung von Wassermolekülen. Bei der katalysierten Hin-Reaktion dienen zwei Moleküle als Substrate, bei der Rück-Reaktion ist nur ein Substratmolekül vorhanden. Der systematische Name wird nach der Konvention „Substrat Gruppe-Lyase“ gebildet. Ist die Rück-Reaktion sehr bedeutsam, kann der Enzymname „Synthase“ enthalten.

EC-Klasse 5: Isomerasen

- Eine Isomerase katalysiert die Änderung der Konformation von Molekülen. Dabei ändert sich die Anzahl der Atome des Moleküls nicht.

EC-Klasse 6: Ligasen

- Ligasen katalysieren die Verbindung zweier Moleküle unter Verbrauch von Energie, die von Triphosphaten, meist ATP, zur Verfügung gestellt wird. „Ligase“ ist im Namen der Enzyme gebräuchlich, in einigen Fällen werden die Namen „Synthase“, „Synthetase“ oder „Carboxylase“ verwendet.

1.5 Homologie: Die Verwandtschaft der Enzyme

Nach der EC-Nomenklatur existieren sechs unterschiedliche Klassen, in die Enzyme eingeteilt werden können. Zu jeder Klasse kennt man derzeit mehrere tausend unterschiedliche Enzymsequenzen [124]. Doch wie ist diese immense Zahl so hochspezialisierter Moleküle entstanden? Eine Analyse der Primärstruktur einiger Sequenzen von Enzymen, die in unterschiedlichen Organismen vorliegen, lässt schnell eine Ähnlichkeit zwischen den

Sequenzen erkennen. Abhängig vom Enzym können die Primärsequenzen sehr ähnlich (homolog) sein.

Diese Ähnlichkeit (Homologie) ist das Resultat einer fortschreitenden Entwicklung, der die Sequenzen (bzw. die codierende DNA) im Laufe der Evolution unterliegen: Die Enzyme haben einen gemeinsamen Ursprung, aus dem alle untersuchten Sequenzen hervorgehen [129]. So sind in den Sequenzen Mutationen entstanden, die sich je nach Organismus, Position und Zeitpunkt des Ereignisses unterscheiden. Einzelne Aminosäuren können durch andere ersetzt worden sein (Substitution), Aminosäuren fallen weg (Deletion, die Sequenz verkürzt sich) oder kommen dazu (Insertion, die Sequenz verlängert sich). Abhängig von den betreffenden Aminosäuren, kann eine Mutation fast folgenlos bleiben oder zu vollständigem Funktionsverlust führen. Entscheidend für die Funktion des gesamten Proteins sind vor allem einzelne Aminosäuren, die im aktiven Zentrum das Substratmolekül binden und primär für die Katalyse verantwortlich sind. Weitere wichtige Positionen in der Sequenz können auch Aminosäuren sein, die besonders für die korrekte dreidimensionale Anordnung des Moleküls verantwortlich sind. Diese Positionen haben sich in den unterschiedlichen Organismen weitgehend erhalten, da sie für die biologische Funktion des Enzyms essentiell sind.

Homologien zwischen Sequenzen sind durch Sequenzalignments identifizierbar. Sind in einem Sequenzalignment mehr als 40% der Aminosäuren der untersuchten Sequenzen identisch, spricht man von einer signifikanten Ähnlichkeit, so dass die Strukturen und/oder die Funktionen der Proteine gleich sind [72, 73, 74]. Eine zufällige Ähnlichkeit durch eine einfache Anordnung der Aminosäuren zu Sequenzen, kann man aufgrund der Vielzahl der Kombinationsmöglichkeiten der Aminosäuren ausschließen. Liegt die Sequenzähnlichkeit unter 40%, spricht man von der *twilight zone.* Sequenzen, die eine Ähnlichkeit in diesem Ausmaß haben, teilen sich nicht mit Sicherheit eine gemeinsame Struktur oder Funktion. Bei der Untersuchung von homologen Sequenzen mittels Alignments, ist aufgrund der Domänenbildung der Proteine (vgl. Abschnitt 1.2) eine Beschränkung der Untersuchung auf einzelne Sequenzabschnitte sinnvoll.

1.6 Methoden zur Sequenzanalyse von Enzymen

Nachdem es gelungen war Enzyme (und Proteine allgemein) zu sequenzieren und Sequenzdatenbanken aufzubauen, war der nächste logische Schritt der Vergleich der neu erhaltenen Sequenzen auf Gemeinsamkeiten. Dazu wurden diese Sequenzen gruppiert und so

lange gegeneinander verschoben, bis das bestmögliche Ergebnis erreicht wurde, d.h. gleiche oder ähnliche Aminosäuren übereinander lagen. Mit Hilfe dieses sogenannten Alignments konnte das Ausmaß der Übereinstimmungen der Sequenzen bestimmt werden. Deren verwandtschaftliches Verhältnis konnte abgeschätzt werden und es gab erste Hinweise auf wichtige Positionen innerhalb der Sequenz. Die Proteine Hämoglobin (Transport von Sauerstoff im Blut) und Cytochrom C (beteiligt am Elektronentransport) gehörten zu den ersten Sequenzen, die von unterschiedlichen Organismen verfügbar waren und auf diese Weise untersucht wurden [130]. Dieses Verfahren war mit wenigen Sequenzen sinnvoll, allerdings können Alignments mit mehreren hundert Sequenzen nicht manuell in adäquater Zeit durchgeführt werden. Die Notwendigkeit, mathematische Algorithmen zu entwickeln, wurde immer größer. Gleichzeitig nahm auch die Rechengeschwindigkeit der Computer immer weiter zu, so dass es sich anbot, den Computer als Werkzeug zur Analyse von biologischen Sequenzen zu nutzen.

Die ersten Algorithmen für die Durchführung automatisierter Alignments stammen von Needleman und Wunsch aus dem Jahr 1970 (“Needleman-Wunsch-Algorithmus“) [131]. Seitdem wurde mit der Etablierung neuer Algorithmen versucht, die Genauigkeit der Alignments zu verbessern und die Suchgeschwindigkeit zu erhöhen [5-13]. Bei den entwickelten Algorithmen, die Sequenzalignments durchführen, unterscheidet man generell Algorithmen für paarweise Alignments (es werden Sequenzen paarweise miteinander verglichen) und multiple Alignments (es werden mehrere Sequenzen miteinander verglichen). Einige Algorithmen untersuchen komplette Sequenzen auf Ähnlichkeit (globales Alignment), andere Algorithmen versuchen die bestmögliche Ähnlichkeit in Sequenzabschnitten zu finden (lokales Alignment). Zu den bekanntesten Suchalgorithmen zählen die Algorithmen von Smith und Waterman (1981) [5] ebenso wie die Algorithmen von Altschul *et al*. (BLAST, 1990) [3], sowie die Algorithmen von Pearson und Lipman (FASTA, 1988) [4].

Derzeit existieren viele Programmen zur Erstellung von Sequenzalignments, die die dargestellten Algorithmen nutzen. Als Beispiele einer Vielzahl der zur Zeit zur Verfügung stehenden Programme seien AMAP [16], MAVID [17], MUSCLE [14, 15] und CLUSTAL W [101, 18] genannt, die globale multiple Alignments erstellen. Das Programm T-Coffee [19, 20] kann zusätzlich auch lokale multiple Alignments erstellen, das Programm BLAST (Basic Local Alignment Search Tool) [3] ist das weltweit am häufigsten genutzte Programm zur Erstellung von lokalen, paarweisen Alignments. Alle Algorithmen und Programme verfolgen das Ziel,

Gemeinsamkeiten zwischen Proteinsequenzen zu erkennen. Sie unterscheiden sich in ihrer Methode, Geschwindigkeit und Präzision.

1.7 Muster von Proteinsequenzen

1.7.1 Der Musterbegriff

In der Literatur wird der Begriff eine Musters im Zusammenhang mit Proteinsequenzen nicht einheitlich benutzt. So werden neben dem klassischen Musterbegriff die Bezeichnungen Motiv, Signatur, Fingerabdruck [1] oder auch Profil verwandt [21]. Alle diese Begriffe bezeichnen aber gemeinsam eine Form, die wesentlichen Eigenschaften der untersuchten Sequenzen darzustellen.

Es existieren kurze, funktionelle Muster, die aus Aminosäurepositionen bestehen, die für die katalytische Aktivität des Enzyms essentiell sind. Strukturmuster geben die Anordnung von Aminosäuren in Proteinstrukturen wieder, z.B. können Muster für Alpha-Helixes oder Beta-Faltblätter erstellt werden [71]. Ein typisches Muster für z.B. ein Beta-Faltblatt ist aber für die Identifizierung einer bestimmten Enzymgruppe mit Hilfe dieses Musters zu generell, da diese Sekundärstruktur in sehr vielen Enzymen vorkommt. Das Muster ist nicht für bestimmte Enzyme spezifisch. Dagegen stellen einige Aminosäuren die dreidimensionale Konformation eines Enzyms sicher, die besonders wichtig bei einer Bindung von Liganden ist bzw. bei der Interaktion von Domänen [54]. Diese Positionen sind typisch für ein Enzym und eignen sich, diese als Fingerabdruck zu verwenden.

Viele in der Literatur beschriebene Sequenzmuster für Enzyme bestehen aus einer Mischung dieser beiden Mustertypen, da in Enzymsequenzen sowohl Aminosäuren konserviert wurden, die für die 3D-Struktur der Aminosäurekette, als auch Aminosäuren, die für die eigentliche Katalyse einer Reaktion verantwortlich sind [2].

1.7.2 Methoden zur Mustererzeugung

Die einfachste aber auch langwierigste Möglichkeit ein Sequenzmuster zu erzeugen, ist die manuelle Erstellung solcher Muster. Auf diese Weise entstehen sehr genaue Muster, da ein Experte Fehler vermeiden kann, die ein Computerprogramm machen könnte. Zusätzlich können spezielle Sequenzen manuell ausgesucht und daraus eine Proteinfamilie gebildet werden, was

zusätzlich die Musterqualität verbessern kann. Neben dieser einfachsten Lösung gibt es eine Vielzahl von Algorithmen und Programmen, die Muster aus alignierten oder nicht-alignierten Sequenzen erzeugen und in unterschiedlichen Formaten ausgeben. Sind Sequenzen nicht aligniert, wird zur Mustererzeugung meist ein Alignment erzeugt, um konservierte Aminosäurepositionen zu identifizieren. Jonassen *et al.* [23] entwickelte z.B. einen Algorithmus, der aus nicht-alignierten Sequenzen Muster erzeugt. Der Algorithmus wurde im Programm PRATT umgesetzt. Ebenso steht ein Server im Internet zur Verfügung, der es Nutzern ermöglicht, interaktiv Sequenzen einzugeben, aus denen Sequenzmuster erzeugt werden [22-24]. Neben PRATT ist der TEIRESIAS Algorithmus etabliert. Dieser erzeugt Sequenzmuster, ohne paarweise Alignments zu erstellen [25, 26]. Eine Vielzahl dieser Algorithmen wurde entwickelt und Programme geschrieben, die ähnlich funktionieren, sich aber darin unterscheiden, auf welchem Weg Muster erzeugt und in welchem Format sie gespeichert werden. Einige Algorithmen begrenzen Muster in ihren Sequenzen oder sie erzeugen Muster mit begrenzter Länge. Andere Algorithmen zeichnen sich durch ihre hohe Rechengeschwindigkeit aus [27-38].

1.7.3 Musterarten und deren Qualitäten

Die in den Abschnitten 1.6 und 1.7.2 dargestellten Algorithmen erzeugen aus zwei oder mehreren Sequenzen Alignments und/oder Sequenzmuster, die die konservierten Aminosäuren der Sequenzen hervorheben. Die Muster sind deterministisch, d.h. ein erstelltes Muster passt auf eine Sequenz oder es passt nicht [47]. Es gibt keinen *score*, der Aussagen über die Signifikanz des Treffers gibt. Die Qualität eines Musters wird dadurch bestimmt, wie gut oder wie schlecht ein Sequenzmuster einer speziellen Proteinfamilie Mitglieder dieser Proteinfamilie identifiziert und gleichzeitig Proteine, die nicht zu dieser Proteinfamilie gehören, von dieser Proteinfamilie unterscheiden kann. Ausschlaggebend für die Qualität des Musters sind die Länge des Musters, die enthaltenen Aminosäuren, die Anzahl von unspezifischen Positionen und das Vorkommen von Lücken.

Ein zweiter Weg zur Darstellung von Mustern, in denen Algorithmen konservierte Positionen innerhalb einer Proteinsequenz speichern, sind probabilistische Muster, d.h. es werden Vektoren oder Matrizen genutzt, um Sequenzpositionen zu speichern [39, 40]. Diese Sequenzmuster werden zum Beispiel definiert durch Profile [43, 44] oder *Hidden Markov Models* (HMMs) [45, 46]. In einigen Fällen werden auch die exakten Grenzen der Muster

hervorgehoben, so dass konservierte Blöcke innerhalb der Sequenzen entstehen [41, 42]. Hier kommen *Position Specific Scoring Matrices* (PSSMs) zum Einsatz [48]. Anders als das deterministische Muster, kann ein probabilistisches Profil oder Modell auch Sequenzen treffen, die nicht exakt mit dem Modell übereinstimmen. Das Ausmaß der Treffer eines Modells wird durch einen Wert (*score*) angegeben. Eine Beurteilung der Qualität eines Treffers ist dadurch möglich. Beide Modelle, das deterministische, sowie das probabilistische, haben ihren praktischen Nutzen. Während die deterministische Spezifikation für den Menschen leicht zu lesen, zu implementieren und zu benutzen ist, sind die Modelle aus dem probabilistischen Ansatz meist sensitiver [47]. Ein Nachteil des probabilistischen Ansatzes ist die hohe CPU Auslastung und damit die Dauer eines Suchvorgangs.

Andere Möglichkeiten zur Mustererzeugung, außer der einfachen Analyse der Primärsequenz, sind alternative Wege, z.B. über die Analyse von Texten und daraus die automatische Extraktion von relevanten Daten (Textmining) [49], der Strukturvorhersage von Proteinen [50, 51], sowie Methoden unter der Berücksichtigung von 3D-Strukturen [52, 53].

1.7.4 Datenbanken für Muster und Proteindomänen

Die bekannteste im Internet frei verfügbare Sammlung von Sequenzmuster und Sequenzfamilien, Profilen, Signaturen, Domänenzuordnungen und weitere Proteineigenschaften ist InterPro [55, 66]. InterPro umfasst die Datenbanken PROSITE [56], die Sequenzmuster und Profile von Proteinen enthält, PRINTS [57], deren Daten auf der *Position Specific Scoring Matrix* Methode (PSSM) basieren, ProDom [58], die das Ergebnis einer automatischen Sequenzclusterung nutzt, sowie Pfam [59], SMART [60], TIGRFAMs [61], PIRSF [62], SUPERFAMILY [63], Gene3D [64] und PANTHER [65], die *Hidden Markov Models* (HMMs) nutzen.
Eine besondere Rolle spielt die PROSITE Datenbank. Die 1988 von Amos Bairoch gegründete Datenbank ist eine der ältesten Datenbanken, die Sequenzmuster speichert und zur Verfügung stellt [67]. Die Einträge werden manuell von Experten generiert und dokumentiert. Die Entscheidungen, aus welchen Gründen bestimmte Sequenzen gruppiert oder spezielle Aminosäuren für ein Muster ausgewählt wurden, sind in der Datenbank dokumentiert. Die Muster der Datenbank sind für Enzyme meist sehr kurz, da sich Naturwissenschaftler, die die Muster erstellen, meist auf Aminosäuren beschränken, die katalytisch aktiv sind. Dennoch haben die Muster eine hohe biologische Signifikanz, so dass unbekannte Mitglieder einer

Sequenzfamilie über diese Muster identifiziert werden können. Dabei besteht die generelle Gefahr, dass mit der Kürze der Muster auch die Signifikanz eingebüßt wird, wenn ein Muster für eine Enzymfamilie eingesetzt wird, das in Enzymen mit unterschiedlicher Funktion vorhanden ist.
Wie andere Datenbanken auch, nutzt PROSITE ein bestimmtes Format zur Speicherung von Sequenzmustern. Aufgrund des Bekanntheitsgrades und der Bedeutung der PROSITE Datenbank, hat sich die von PROSITE genutzte Syntax durchgesetzt, so dass Forschergruppen meist dieses Format zur Darstellung von konservierten Aminosäuren von Proteinsequenzen verwenden.

1.7.5 Die Notwendigkeit, Muster zu erstellen

Die Erzeugung von Sequenzmustern ist wichtig, wenn die Struktur oder die Funktion eines Proteins analysiert werden soll, da ein Muster wenige bedeutende Aminosäuren aus einer Sequenz hervorhebt. Ob es sich dabei um ein Strukturprotein oder Enzym handelt, spielt keine Rolle. Ein Sequenzmuster kann kurz und präzise den prinzipiellen Aufbau von zum Beispiel Kollagen oder Seidenfibroin vermitteln, aber gerade bei Enzymen hat sich gezeigt, dass speziell die bei der Katalyse direkt beteiligten Aminosäuren besonders konserviert sind [68], so dass die Erstellung von Sequenzmustern besonders bei Enzymsequenzen erfolgsversprechend ist. Bei Enzymen sind Sequenzmuster aufgrund ihrer extremen Vielfalt in Aufbau und Funktion von großer Bedeutung. Enzyme können aktive oder allosterische Zentren besitzen und vielfältige Reaktionen katalysieren. Zum Verständnis von z.B. Funktion und Aufbau der Enzyme, der Einteilung von unbekannten Enzymen in Proteinfamilien, der Identifizierung von phylogenetischen Verwandtschaftsverhältnissen oder bei der Zuweisung von Funktionen nicht näher definierten Sequenzen, sind Sequenzmuster unerlässlich. Aus diesem Grund ist es sinnvoll Musterdatenbanken zu erstellen, in denen das Wissen über Enzymsequenzen gespeichert und das bei der Klassifizierung einer Sequenz schnell zur Verfügung gestellt wird.

1.8 Clusteranalyse von biologischen Sequenzen

Durch die Weiterentwicklung von experimentellen Sequenziertechniken in der molekularen Biologie, sowie durch eine Vielzahl von Genomprojekten, unterliegen öffentliche Proteindatenbanken einem so rapiden Wachstum, dass es immer schwieriger sein wird, die erhaltene Datenmenge zu analysieren. Eine Clusterung dieser Daten in sinnvolle Gruppen ist ein häufig genutztes Verfahren in der Bioinformatik [89, 90]. Oftmals werden Sequenzen nach ihrer Ähnlichkeit geclustert, die durch Computerprogramme wie BLAST oder FASTA errechnet werden [4, 79, 91]. Das Ergebnis der Clusterung kann genutzt werden, um, basierend auf Sequenzähnlichkeit, Proteinfamilien zu bilden. Aber gerade die Identifizierung und Bestimmung von Sequenzfamilien stellt ein Problem dar.

Eines der am häufigsten genutzten Clusterverfahren ist das *single linkage clustering* [133]. Es handelt sich hierbei um eine hierarchische Clusterung, bei dem Cluster schrittweise vereinigt werden. Eingesetzt in der Bioinformatik, ist das *single linkage* Verfahren dafür bekannt, biologisch sinnvolle Gruppen von homologen Sequenzen bilden zu können. Dabei muss das Problem der Domänenstruktur der Proteine beachtet werden, denn die meisten Proteine sind modular aus Domänen aufgebaut (vgl. Abschnitt 1.2). Vergleicht man verschiedene Sequenzen miteinander, ist es daher sinnvoll, sich bei der Untersuchung von Sequenzen, basierend auf Sequenzähnlichkeit, auf Subsequenzen, die Domänen, zu konzentrieren, um ein sinnvolles Ergebnis zu erhalten. Das Ergebnis der Clusterung von Enzymsequenzen kann dazu genutzt werden, die EC-Klassifizierung der Enzyme zu prüfen, indem Gemeinsamkeiten oder Unterschiede zwischen geclusterten Enzymsequenzen mit unterschiedlichen EC-Nummern untersucht werden.

1.9 Reaktionsmatrizen zur Untersuchung von Enzymreaktionen

Reaktionsmatrizen (R-Matrizen) sind mathematische Operatoren, die Elektronentransfermuster von chemischen Reaktionen speichern. In ihnen sind Informationen über Atome im Molekül enthalten und sie geben Auskunft, welche Bindungen zwischen den Atomen gebildet oder gespalten werden. Diese Matrizen können dazu genutzt werden, chemische Reaktionen, z.B. Reaktionen, die von Enzymen katalysiert werden, miteinander zu vergleichen. Für die Erstellung einer R-Matrix ist allerdings eine genaue Zuordnung von Atomen der Edukte und Produkte der Reaktionen nötig. Diese Atomzuordnungen können manuell erfolgen, es existieren aber auch Algorithmen, die Edukt- und Produktatome automatisch erkennen und

diese im Edukt- und Produktmolekül markieren. Effektive Algorithmen, die Moleküle miteinander vergleichen können und aus der Graphentheorie stammen, sind die sogenannten MCS-Algorithmen, wobei MCS eine Abkürzung für „Maximal Common Subgraph“ ist. Viele Forschergruppen nutzen die Möglichkeit, mittels MCS-Algorithmen Moleküle miteinander zu vergleichen [83-85]. Allerdings haben viele MCS-Algorithmen Nachteile. So dauert eine Berechnung umso länger, je größer die zu vergleichenden Moleküle sind. Zudem sind viele existierende Algorithmen zu unflexibel, um komplexe Veränderungen zwischen Molekülen zu erkennen [86].

Markus Leber hat in seiner Dissertation einen c-MCS Algorithmus entwickelt, der die oben beschriebenen Nachteile weitgehend tilgt. Dieser sehr flexible und schnelle Algorithmus basiert auf einer Kombination aus einer Variante des Bron-Kerbusch Algorithmus‘ [87] und dem McGregor Algorithmus [84]. Der spezielle c-MCS Algorithmus wurde in seiner Arbeit dazu genutzt, die größte gemeinsame Teilstruktur von unterschiedlichen Molekülen zu identifizieren und mittels erstellter Reaktionsmatrizen enzymatische Reaktionen automatisch miteinander zu vergleichen. Die Auswertung dieser Ergebnisse ist eine sinnvolle Ergänzung, um die Ergebnisse der Clusteranalyse, die auf Sequenzähnlichkeit beruht, zu überprüfen.

1.10 Zielsetzung

In der heutigen Zeit werden sehr viele unbekannte Proteine entdeckt und sequenziert, so dass Sequenzdatenbanken mit großer Geschwindigkeit wachsen. Eine manuelle, experimentelle Einschätzung zur Einteilung der Sequenzen in Familien oder die Untersuchung der Struktur und Funktion von Enzymen ist sehr langwierig und bei der Menge der zur Verfügung gestellten Daten praktisch nicht möglich. Wie die Erfahrung der letzten Jahre zeigte, wird in Zukunft die Geschwindigkeit des Wachstums von Sequenzdatenbanken noch zunehmen. Eine möglichst verlässliche, automatische Methode zur Einteilung von Proteinen in Familien, die Identifizierung von Proteindomänen und die automatische Identifizierung von konservierten Aminosäuren ist daher zwingend erforderlich, um bereits entdeckte Proteine zu klassifizieren und die Funktion von neuen Sequenzen schnell bestimmen zu können.

Sequenzmuster sind dabei ein essentielles Werkzeug, die kurz und präzise, konservierte Aminosäuren in Proteinen darstellen. Manuell erzeugte Sequenzmuster, wie sie seit Jahren von Experten für die Datenbank PROSITE erstellt werden, haben den Nutzen von Sequenzmuster gezeigt.

Die Clusteranalyse ist in der Biologie eine weit verbreitete und etablierte Methode, Sequenzen mit Hilfe des Computers automatisch zu gruppieren. Einige Sequenzdatenbanken basieren bereits auf der Idee, Proteinfamilien auf diese Weise automatisch zu erstellen und genauer zu untersuchen.

In der vorliegenden Arbeit werden alle derzeit bekannten Enzymsequenzen, die eine vollständige EC-Nummer tragen, analysiert. Basierend auf der Annahme, dass homologe Sequenzdomänen geclustert werden, können durch eine Untersuchung der Sequenzzusammensetzungen der Cluster die evolutionären bzw. funktionellen Beziehungen unterschiedlicher EC-Nummern untereinander dargestellt werden. Konservierte Aminosäuren, die sich über unterschiedliche EC-Nummern hinweg erhalten haben, sind wahrscheinlich für die Struktur oder Funktion der Enzyme von besonderer Bedeutung. Für diese Analyse werden aus den Sequenzen bestimmter Cluster Sequenzmuster erstellt, die für die EC-Nummern der im Cluster enthaltenen Sequenzen typisch sind. Auf diese Weise soll eine Datenbank entstehen, die für alle vollständigen EC-Nummern mindestens ein Sequenzmuster enthält. Die Datenbank soll folgende Eigenschaften erfüllen:

- Vollständigkeit: Möglichst alle existierenden Muster sollen entdeckt werden.
- Eindeutigkeit: Die Grenzen der Muster werden aufgrund der Domänenlänge limitiert.
- Spezifität: Die Muster sollen so spezifisch und sensitiv wie möglich sein.
- Dokumentation: Informationen, aus welchen Sequenzen und Clustern die Muster erstellt wurden, werden geliefert.
- Aktualität: Die Datenbank soll in regelmäßigen Abständen aktualisiert werden.
- Schnelligkeit: Alle Daten sollen in einer hohen Geschwindigkeit zur Verfügung stehen.

Die Beziehung zwischen Sequenzähnlichkeit und Funktion von Enzymen soll zwischen Sequenzen, die sich in einem Sequenzcluster befinden und unterschiedliche EC-Nummern tragen, mittels Reaktionsmatrizen untersucht werden, so dass eine Aussage möglich ist, in welchem Umfang homologe Enzymsequenzen sich ähnliche Reaktionsmechanismen teilen. Ob ähnliche Moleküle umgesetzt werden, soll eine Analyse klären, die zu diesem Zwecke durchgeführt wird. Dabei wird die größte gemeinsame Teilstruktur, der bei den von Enzymen katalysierten, chemischen Reaktionen beteiligten Moleküle, mittels des c-MCS Algorithmus‘ ermittelt.

In welchem Maße katalytische Mechanismen während der Evolution konserviert wurden und welche Aussagekraft die erstellten Muster haben, wird anhand von Beispielen mit Hilfe von 3D-Darstellungen der betreffenden Sequenzstrukturen und der Vergleich mit möglichen Mustern der PROSITE-Datenbank untersucht.

2 Daten, Algorithmen und Methoden

2.1 Übersicht

Die folgende Übersicht zeigt den inhaltlichen Verlauf und die Strategie der Arbeit.

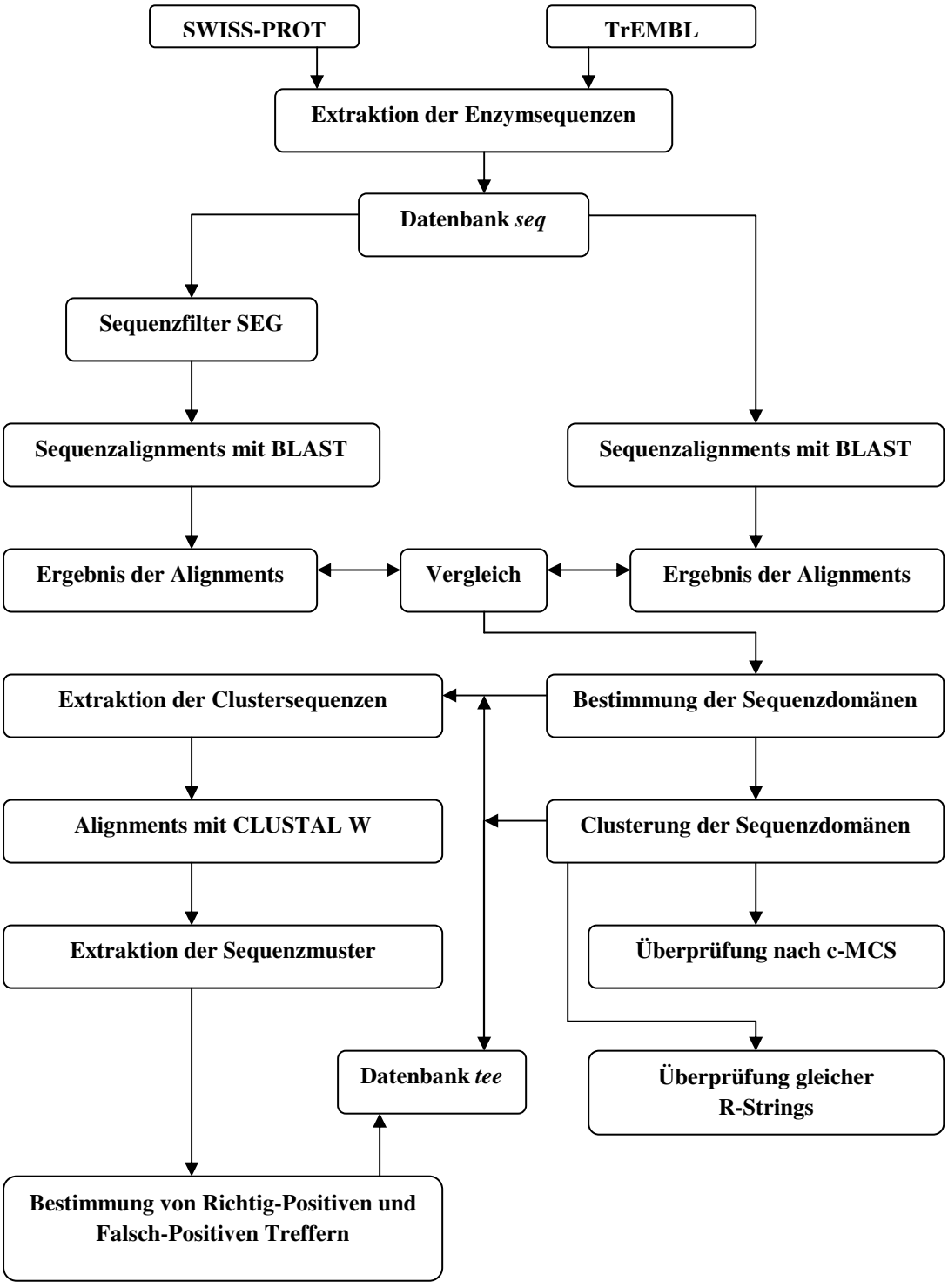


Abbildung 2-1: Graphische Darstellung des Arbeitsverlaufs.

2.2 Strategie und Verlauf der Arbeit

Ziel der Arbeit ist es, mit Hilfe einer Clusteranalyse homologe Enzymsequenzen zu gewinnen und eine Identifizierung, Generierung und Analyse von konservierten Sequenzmustern automatisiert durchzuführen. Zudem sollen die Edukte, Produkte und Co-Substrate von chemischen Reaktionen unterschiedlicher Enzyme, die bei verschiedenen E-Werten geclustert wurden, mittels des c-MCS Algorithmus‘ [86] auf ihre größte gemeinsame Teilstruktur untersucht werden. Der Vergleich der Clusterung von Enzymsequenzen aufgrund von Sequenzähnlichkeit mit der Clusterung von Enzymen aufgrund von gleichen R-Strings soll unter Berücksichtigung der EC-Klassifizierung Aufschluss über eine Beziehung von Sequenzähnlichkeit und Reaktionsmechanismus liefern.

Im ersten Schritt der Arbeit werden aus den Proteindatenbanken SWISS-PROT (vgl. Abschnitt 2.11.1) und TrEMBL (vgl. Abschnitt 2.11.2) alle Enzymsequenzen extrahiert, die mindestens eine vollständige EC-Nummer erhalten haben. Aus diesen Informationen wird die Datenbank *seq* erstellt, die aus Tabellen besteht, die neben Sequenzen und EC-Nummern auch zusätzliche Informationen über Eigenschaften der Sequenzen liefern, sofern diese in den Ausgangsdatenbanken enthalten sind. Anschließend werden aus allen in *seq* enthaltenen Sequenzen zwei vollständige all-vs-all BLAST Alignments durchgeführt. Während im zweiten Alignmentdurchgang alle Sequenzen unverändert aus der Datenbank entnommen werden, wird im ersten Durchgang der Sequenzfilter SEG [94, 95] dazu genutzt, Sequenzen mit geringer kompositorischer Komplexität, aus den Alignments auszuschließen. Das Resultat aus beiden Durchgängen wird in der Datenbank *seq* gespeichert. Ein Vergleich beider Alignmentergebnisse soll die Qualität des Datensatzes verbessern. Der E-Wert, der bei der Erstellung eines Alignments bestimmt wird (vgl. Abschnitt 2.3.1) dient dazu, die Verlässlichkeit eines Alignments zu bestimmen. Der maximale E-Wert, den ein Alignment erreichen darf, wird auf 10^{-2} festgelegt. Wird dieser Wert überschritten oder unterschreitet die Sequenzlänge eine Mindestlänge von 60 Aminosäuren, werden diese Alignments nicht in die Datenbank aufgenommen (vgl. Abschnitt 2.4.2).

Die erhaltenen Sequenzdomänen werden im nächsten Schritt nach dem *single linkage* Verfahren geclustert. Die Anzahl der Sequenzabschnitte, die durch Alignments erhalten wurden, wird durch bestimmte Verfahren reduziert. So werden z.B. Sequenzabschnitte, die innerhalb definierter Grenzen starten oder enden, zu einem Abschnitt mit eindeutigen Grenzen zusammengefasst. Die Clusterung wird mit den ähnlichsten Sequenzen begonnen. Im Verlauf der Clusterung werden immer unähnlichere Sequenzen geclustert, so dass die Anzahl der

Cluster abnimmt, die Anzahl der Sequenzen innerhalb der Cluster zunimmt. Damit entstehen Clusterbäume homologer Sequenzen. Als Ähnlichkeitsmaß für die Clusterung dient der bei den lokalen Alignments bestimmte E-Wert.
Aus bestimmten Clustern werden Sequenzmuster erstellt. Alle Sequenzen eines Clusters werden einem globalen Alignment mit dem Programm CLUSTAL W unterzogen. Aus diesem Alignment wird das Muster im PROSITE-Format extrahiert. Alle Muster, sowie die Ergebnisse der Suche nach Richtig-Positiven und Falsch-Positiven Treffern dieser Muster in den Datenbanken SWISS-PROT und TrEMBL werden in der Datenbank *tee* gespeichert. Diese Muster sollen darüber Aufschluss geben, in welchem Maße diese Muster dazu dienlich sind, neue Sequenzen, die nicht klassifiziert wurden, mit Hilfe dieser Muster zu identifizieren.

Bei den E-Werten 10^{-2}, 10^{-40}, 10^{-80}, 10^{-120}, 10^{-160} und 10^{-181} werden Sequenzen, die verschiedene EC-Nummern tragen und sich bei den angegebenen E-Werten in einem Cluster befinden, analysiert. Mit Hilfe des c-MCS Algorithmus‘ werden die Edukte, Produkte und Co-Substrate der chemischen Reaktionen, die von Enzymen katalysiert werden und deren Sequenzen sich in einem Cluster befinden, auf ihre größte gemeinsame Teilstruktur untersucht. Zusätzlich wird die durch Sequenzähnlichkeit erhaltenen Clusterung mit der Clusterung der Enzyme anhand identischer R-Strings verglichen. Das Ergebnis beider Analysen soll darüber Aufschluss geben, in welchem Maße Sequenzähnlichkeit und Reaktionsmechanismus korrelieren.

Anhand von Beispielen wird stellvertretend für den kompletten Datensatz auf die Ergebnisse der beschriebenen Verfahrensweisen eingegangen.

2.3 Beschreibung von BLAST und die Berechnung des E-Werts

2.3.1 Der BLAST Algorithmus

Ein Sequenzalignment dient dem Vergleich zweier (paarweises Alignment) oder mehrerer (multiples Alignment) Strings. Das Alignment, im Deutschen oft in Anlehnung an den englischen Begriff „Alignierung“ genannt, ist eine in der Bioinformatik oft verwandte Methode, um eine evolutionäre und funktionelle Verwandtschaft, sog. Homologie, von Nucleotid- oder Aminosäuresequenzen zu identifizieren [3]. Bei einem Alignment werden Sequenzen gruppiert und solange gegeneinander verschoben, bis gleiche oder ähnliche Elemente der verglichenen Sequenzen übereinander liegen. Bei kleinen Datensätzen ist es möglich, ein Alignment manuell auszuführen. Je länger die zu vergleichenden Sequenzen sind, bzw. je mehr Sequenzen verglichen werden sollen, desto komplizierter wird ein solches manuelles Vorgehen werden. Es wurden daher Algorithmen entwickelt, die Alignments automatisiert durchführen können [4].
Eines dieser Programme ist das in dieser Arbeit genutzte BLAST Programm. BLAST ist eine Abkürzung für *Basic Local Alignment Search Tool* und bezeichnet eine Sammlung der weltweit am meisten genutzten Programme zur Analyse von biologischen Sequenzen. Es wurde im Jahr 1990 von Altschul *et al.* veröffentlicht [3] und zwischenzeitlich modifiziert [96]. Das Programm BLASTP (P deutet im Namen darauf hin, dass Proteinsequenzen genutzt werden) verarbeitet Proteinsequenzen und vergleicht diese gegen eine Sequenzdatenbank. Diese Datenbank, gegen die gesucht wird, kann lokal erstellt werden.
Am 01.04.2008 wurde die BLAST Version 2.2.18 veröffentlicht. In dieser Arbeit wurde die Version 2.2.16 eingesetzt. Bei den Berechnungen wurde die *Blocks Substitution Matrix 62*, BLOSUM62 verwendet [97]. Die folgenden Abschnitte geben einen Überblick über das Funktionsprinzip von BLAST und erklären die Bedeutung der errechneten Werte.

Der BLAST Algorithmus nutzt eine Substitutionsmatrix, die bei einem Vergleich von zwei Sequenzen für jedes Aminosäurepaar i und j den *score* s_{ij} definiert. Liegt bei einer Sequenzpaarung in einem Abschnitt mindestens ein *hit* (Treffer) vor und liegt der *score* über einem zuvor bestimmten Grenzwert, liegt ein *High-Scoring Segment Pair* (*HSP*) vor. Die folgenden vier Werte werden von BLAST ausgegeben. Mit Hilfe dieser Werte [3, 80, 96] ist eine Einschätzung der Signifikanz des Sequenzvergleichs möglich:

***identity*:** Die Identität ist eine prozentuale Angabe, die aussagt, wie viele Aminosäuren der Datenbanksequenz mit Aminosäuren der Suchsequenz übereinstimmen.

***raw score*:** Der *raw score* kann nur dann sinnvoll interpretiert werden, wenn detailierte Kenntnisse über das eingesetzte Scoring System vorhanden sind, da bei der Berechnung des *raw scores* statistische Parameter eingehen. Der *raw score* ist die Summe der zuvor ermittelten *scores*, von denen noch die Werte für Insertionen und Deletionen abgezogen werden.

***bit score*:** Nach der Theorie der mathematischen Statistik wird angenommen, dass in einem einfachen Protein an jeder Position mit der gleichen Wahrscheinlichkeit jede Aminosäure vorhanden sein kann. Der *score* für zwei zufällige Aminosäuren sollte nach der folgenden Formel negativ sein:

$$\sum_{ij} P_i P_j s_{ij}$$

Bei den gegebenen Parametern P_i und s_{ij} besagt die Grundtheorie, dass mit Hilfe der beiden Parameter λ und K der *HSP score* in einen normalisierten *score* umgewandelt werden kann [98]. Die Einheit dieses Wertes ist *bit* [99]. Der normalisierte *score S'* wird durch folgende Formel definiert:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

***E-value (E-Wert)*:** Wenn zwei Proteine mit zufälliger Aminosäurezusammensetzung der Längen n und m miteinander verglichen werden, wird die Anzahl E unterschiedlicher *HSPs* mit normalisiertem *score* von mindestens S' erwartet, die zufällig entsteht. $N = mn$ ist dabei der Suchraum. Der Wert E wird mit folgender Formel berechnet:

$$E = \frac{N}{2^{S'}}$$

Bei einer typischen Datenbanksuche, bei der ein Protein von einer Länge von 250 Positionen gegen eine Datenbank verglichen wird, deren Sequenzen aus 50 Millionen Aminosäuren

bestehen, würde ein normalisierter *score* von ~35 bit nötig sein, um einen signifikanten E-Wert von 0,05 zu erzielen [96].
Einfach ausgedrückt sagt der E-Wert eines Alignments aus, mit welcher Wahrscheinlichkeit das erzielte Alignment zustande käme, wenn zwei zufällige Sequenzen miteinander verglichen werden, wenn Parameter, wie Datenbankgröße und enthaltene Aminosäuren, konstant sind. Je kleiner der E-Wert eines Alignments ist, desto signifikanter ist das Alignment. Der E-Wert ist somit ein Wert, mit dessen Hilfe sich direkt die Qualität eines Alignments einschätzen lässt. Dem E-Wert kommt während der Clusteranalyse von Proteinsequenzen eine besondere Bedeutung zu, da nach diesem Wert die Ähnlichkeit zwischen Sequenzen und die Zusammensetzung der Cluster bestimmt wird.

2.3.2 Einschätzung der Signifikanz von Alignments bei verschiedenen E-Werten

Aus dem vorherigen Abschnitt geht hervor, dass je kleiner der E-Wert eines Alignments ist, desto signifikanter ist das Alignment. *The Southwest Biotechnology and Informatics Center* (SWBIC) [100] schätzt die Bedeutung verschiedener E-Werte wie folgt ein:

Bei E-Werten unter 10^{-50} deutet ein *hit* (Treffer zur Vergleichssequenz) auf einen Vergleich sehr ähnlicher Sequenzen hin. Die Wahrscheinlichkeit, dass die untersuchten Sequenzen evolutionär verwandt sind, ist sehr groß. Es empfiehlt sich, die Sequenzen genauer, z.B. mit weiteren Programmen, zu untersuchen.

Liegen die E-Werte im Bereich zwischen 10^{-50} und 10^{-2} hat der *hit* Ähnlichkeit zur Vergleichssequenz. Eine evolutionäre Verwandtschaft ist möglich, weitere Untersuchungen sind anzuraten. Wahrscheinlich gehören die untersuchten Sequenzen zu einer gemeinsamen Sequenzfamilie.

Ist der E-Wert größer als 10^{-2}, hat der *hit* eine sehr schwache Ähnlichkeit mit der untersuchten Sequenz. Wird eine evolutionäre Verwandtschaft vermutet, ist sie sehr schwach und/oder liegt aus evolutionärer Sicht sehr lange zurück. Sind diese Hintergrundinformationen vorhanden, empfiehlt sich eine weitere Analyse.

Bei Werten von eins und darüber, besteht eine sehr geringe Ähnlichkeit. Die Sequenzen gehören mit großer Wahrscheinlichkeit keiner gemeinsamen Sequenzfamilie an.

2.4 Clusteranalyse von homologen Enzymsequenzen

2.4.1 Theorie der Clusterung

Dieser Abschnitt beschäftigt sich mit der grundlegenden, theoretischen Idee der Clusteranalyse. Das in dieser Arbeit verwendete Verfahren der Clusterung wird in der Literatur als hierarchisch, agglomerativ bezeichnet. Demnach existieren zu Beginn der Clusterung sehr viele Gruppen, die jeweils nur ein Element enthalten. Im nächsten Schritt werden die Elemente miteinander verglichen und ähnliche Elemente werden zu Gruppen (Cluster) zusammengefasst. Innerhalb einer Gruppe sind die enthaltenen Elemente sehr ähnlich, die Ähnlichkeit zwischen den Gruppen ist besonders groß. Im Verlauf der Clusterung werden die Elemente unterschiedlicher Gruppen fortwährend miteinander verglichen. Dabei wird das Beurteilungsmaß, welche Elemente als „ähnlich“ gelten, schrittweise herabgesetzt. In weiteren Stufen der Clusterung werden Elemente, die auf der vorherigen Stufe nicht geclustert wurden, in diesem Schritt als „ähnlich“ eingestuft und nun geclustert. Sind die Elemente in dieser herabgesetzten Beurteilung zwischen Clustern ähnlich, fusionieren die Elemente der Cluster zu einem Cluster, das aus allen Elementen beider ursprünglicher Cluster besteht. Auf diese Weise kondensieren sich alle Cluster so lange, bis an der Spitze ein Cluster mit allen Elementen entsteht. Nach diesem Verfahren entsteht ein hierarchischer Clusterbaum. Eine besondere Bedeutung über den Erfolg und der Aussagekraft der Clusterung von biologischen Sequenzen, bzw. von Elementen aller Art, hat ein adäquates Ähnlichkeitsmaß, das zuverlässig Informationen zu der Ähnlichkeit der Elemente liefert. Neben einem Ähnlichkeitsmaß wird für die Clusterung zusätzlich ein Distanzmaß benötigt. Dieses Maß ist nicht eindeutig, wenn ein Cluster aus mehreren Elementen besteht. Die Auswahl, welches Maß die Distanz zwischen Elementen bestimmt, ist für das Ergebnis der Clusterung besonders wichtig. Das Distanzmaß wird anhand des zu erwartenden Resultats und anhand der Elemente, die geclustert werden, gewählt [80].

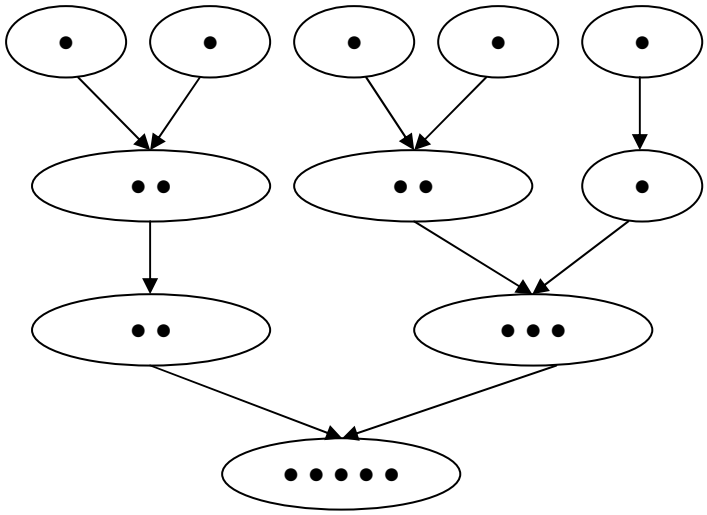


Abbildung 2-2: Im hierarchischen, agglomerativen Clusterverfahren werden Elemente schrittweise geclustert. Ein Clusterbaum entsteht.

2.4.2 Praktische Clusteranalyse von Enzymsequenzen

Die in dieser Arbeit verwendete Clusteranalyse und die Bestimmung der Domänenstruktur geht auf die Arbeit von Christian aus dem Spring zurück [80]. Die folgenden Abschnitte fassen das grundlegende Vorgehen bei der Clusteranalyse und bei der Bestimmung der Domänengrenzen nach seiner Idee zusammen.

Zu Beginn des Verfahrens werden allen aus SWISS-PROT und TrEMBL extrahierten Enzymsequenzen zwei all-vs-all BLAST Alignments unterzogen. Im ersten all-vs-all Alignment wird der in BLAST enthaltene Sequenzfilter SEG [94, 95] genutzt, um Sequenzen mit geringer kompositorischer Komplexität, aus dem Alignment auszuschließen. Im zweiten Alignment wird dieser Filter nicht genutzt. Die hier eingegebenen Sequenzen werden vollständig aus SWISS-PROT und TrEMBL übernommen. Das Ergebnis aus beiden all-vs-all Alignments wird nach Abschluss der Berechnungen miteinander verglichen. Auf diese Weise wird ein Datensatz gewonnen, der Sequenzalignments enthält, deren Sequenzähnlichkeit nicht auf Sequenzbereiche geringer kompositorischer Komplexität beruht.

Um die Qualität der Sequenzvergleiche weiter zu verbessern, werden Sequenzpaarungen aus dem Datensatz entfernt, die

1. nicht in beiden all-vs-all Alignments vorkommen.
2. einen größeren E-Wert als 10^{-2} erzielen.
3. eine Mindestlänge von 60 Aminosäuren nicht erreichen.
4. in der Beschreibung der Sequenz die Wörter „Probable“, „Putative“, „Hypothetical“, „Possible“ oder „Potential“ enthalten.

2.4.2.1 Identifzierung von Proteindomänen

Die Grundlage der Identifizierung von Proteindomänen bildet das Ergebnis der all-vs-all BLAST Sequenzalignments. Anhand dieses Resultats wird versucht, die Grenzen der Proteindomänen zu bestimmen. Dies geschieht in drei Schritten.
Im ersten Schritt werden alle Alignments bestimmt, die für eine Sequenz verfügbar sind. Um eine Häufung von Start- und Endpositionen aller Alignments zu identifizieren, werden alle Alignments übereinandergelegt. Die daraus resultierenden Werte von Start- und Endpositionen lassen sich auf diese Weise ermitteln. In der Abbildung 2-3 wird dieser Vorgang illustriert. Die Höhe der Ausschläge ist äquivalent mit der Häufung von Start- und Endregionen.

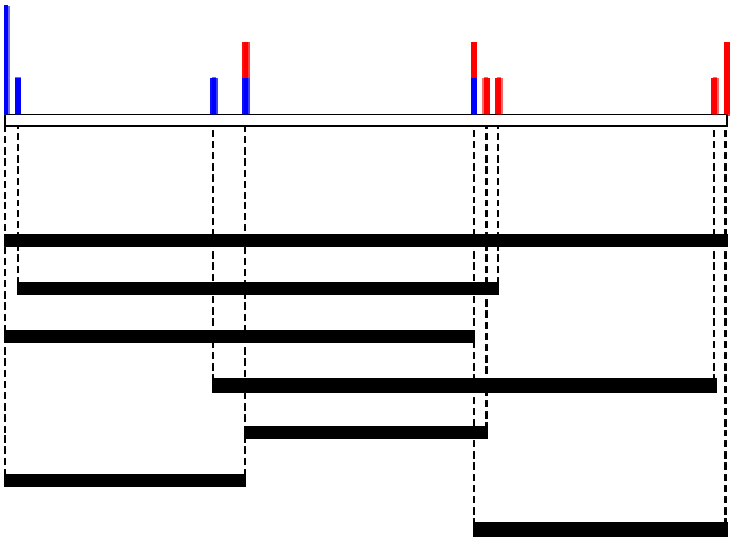


Abbildung 2-3: Alle mittels Alignments erhaltenen Abschnitte werden an die Sequenz angelegt, von der die Abschnitte stammen. Die resultierende Häufung der Start- und Endpositionen bilden die Grundlage zur Bestimmung der Proteindomänen, nach aus dem Spring [80].

Um aus der Häufung der Start- und Endpositionen der Alignments eine eindeutige Domänengrenze zu bestimmen, werden die N- und C-Termini der Alignments unabhängig voneinander untersucht. Ein Raster mit einer Fenstergröße von 15 Aminosäuren wird über die Sequenz gelegt und für alle N- sowie für alle C-Termini, die innerhalb dieser 15 Aminosäuren liegen, schrittweise eindeutige Domänengrenzen bestimmt. Grundlage hierfür ist die Anzahl der Alignments, die an betreffender Position starten bzw. enden. Existieren mehrere Häufungen gleicher Dimension innerhalb des Bereichs der 15 Aminosäuren, werden zusätzliche Informationen, d.h. die angrenzenden Start- oder Endpositionen, für eine Entscheidungsfindung genutzt. Ist das Ergebnis auch nach diesem Schritt gleich, wird jeweils links und rechts der Position nach Maxima gesucht, bis ein eindeutiges Ergebnis vorliegt. Wird auf diese Weise das Ende der Sequenz erreicht, wird das Maximum als Domänengrenze gewählt, das am nächsten zum Sequenzende liegt.

Im letzten Schritt zur Bestimmung der Domänenstruktur werden, anders als im Schritt zuvor, die N- und C-Termini nicht unabhängig voneinander untersucht, um Domänen mit unterschiedlichen Längen zu berücksichtigen. Dabei sollen zuvor bestimmte Domänenregionen nicht zu einer großen Region zusammengefasst werden. Das wird dadurch erreicht, indem alle ähnlichen Domänen, die im gleichen Bereich liegen, mit Hilfe einer Clusteranalyse, gruppiert werden. Eine detailierte Übersicht dieses Verfahrens liefert die folgende Abbildung 2-4.

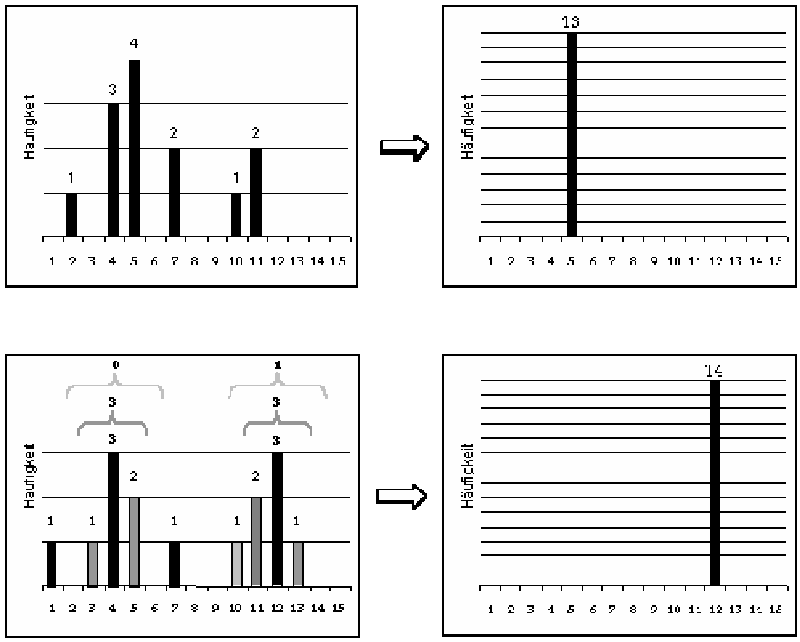


Abbildung 2-4: Detailierte Darstellung zur Identifizierung der Domänengrenzen. Im Bereich von 15 Aminosäuren wird die Aminosäureposition als Domänengrenze bestimmt, an der die meisten Alignments starten. Liegen in diesem Bereich zwei Maxima, werden die umgebenen Startpositionen der Maxima dazu herangezogen, die Domänengrenze eindeutig zu bestimmen, nach aus dem Spring [80].

2.4.2.2 Sequenzclusterung zur Ermittlung der Domänengrenzen

Um mittels Sequenzalignments Domänengrenzen eindeutig zu bestimmen, werden die bisherigen Verfahren zur Ermittlung der Grenzen weiter verfeinert (vgl. Abschnitt 2.4.2.1). Zu diesem Zweck werden die bisher gewonnenen Sequenzabschnitte nach dem *complete linkage* Verfahren geclustert (vgl. Abbildung 2-5).

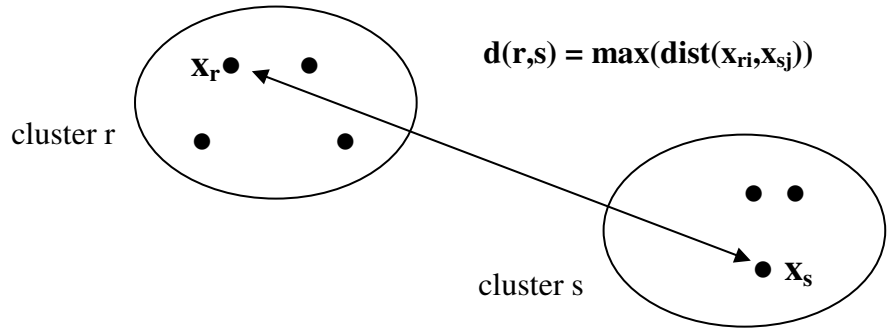


Abbildung 2-5: Das Distanzmaß zwischen zwei Clustern nach dem *complete linkage* Verfahren, nach aus dem Spring [80].

Es werden alle durch Alignments gewonnenen Abschnitte, die im gleichen Sequenzbereich liegen, miteinander verglichen. Je größer die Übereinstimmungen sind, desto kleiner sind die nicht überlappenden Grenzen an den Enden der Abschnitte. Nun werden alle Abschnitte, die in der gleichen Sequenzregion liegen geclustert, indem im ersten Schritt alle Abschnitte verwendet werden, die untereinander sehr ähnlich sind. Es werden nur Abschnitte gruppiert, die sich in den N- und C-Termini um bis zu 45 Aminosäuren unterscheiden, um gewährleisten zu können, dass nur ähnliche Abschnitte berücksichtigt werden. Im folgenden Schritt werden alle Sequenzabschnitte einer Sequenz, die in einem Cluster enthalten sind, durch eine eindeutige Domänengrenze ersetzt. Dieses Verfahren wird mit allen Regionen durchgeführt, um alle Domänengrenzen eindeutig bestimmen zu können.

2.4.3 Clusterung von homologen Enzymdomänen

Die durch Sequenzalignments erhaltenen homologen Sequenzabschnitte werden nach dem *single linkage* Verfahren geclustert. Ein Überblick über die Vorgehensweise bei diesem Verfahren wird mit Abbildung 2-6 geliefert.

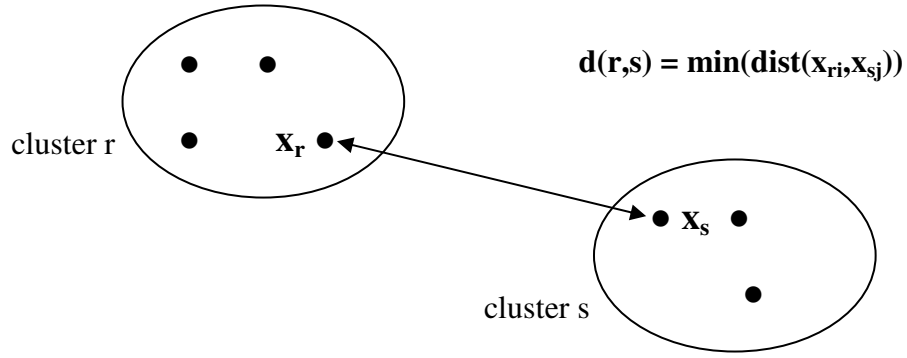


Abbildung 2-6: Das Distanzmaß zwischen zwei Clustern nach dem *single linkage* Verfahren, nach aus dem Spring [80].

Die Elemente der Cluster bilden die nach lokalen all-vs-all BLASTP Alignments erhaltenen Sequenzabschnitte, deren Grenzen in mehreren Schritten verfeinert werden. Auf diese Weise wird auch die Anzahl der Regionen reduziert. Nach der Clustertheorie des *single linkage* Verfahrens (vgl. Abschnitt 2.4.1) werden zunächst alle Sequenzen in jeweils ein Cluster gruppiert, wobei mit den Sequenzen begonnen wird, die untereinander sehr ähnlich sind. Das Maß für die Ähnlichkeit ist der beim Alignment erhaltene E-Wert, der von BLASTP ermittelt wird. Bei steigendem E-Wert werden alle Cluster fusioniert, wenn zwischen mindestens einem Element eines Clusters mit mindestens einem Element eines anderen Clusters ein Alignment mit jeweiligem E-Wert vorhanden ist. Mit steigendem E-Wert werden dadurch die Cluster des Clusterbaums immer größer, jeweils neue Elemente der Cluster werden untereinander unähnlicher.

In dieser Arbeit werden Sequenzen in 180 Schritten geclustert, da E-Werte von Alignments zwischen 10^{-181} bis 10^{-2} berücksichtigt werden.

2.5 Mustererstellung

2.5.1 Clusterauswahl zur Mustererstellung

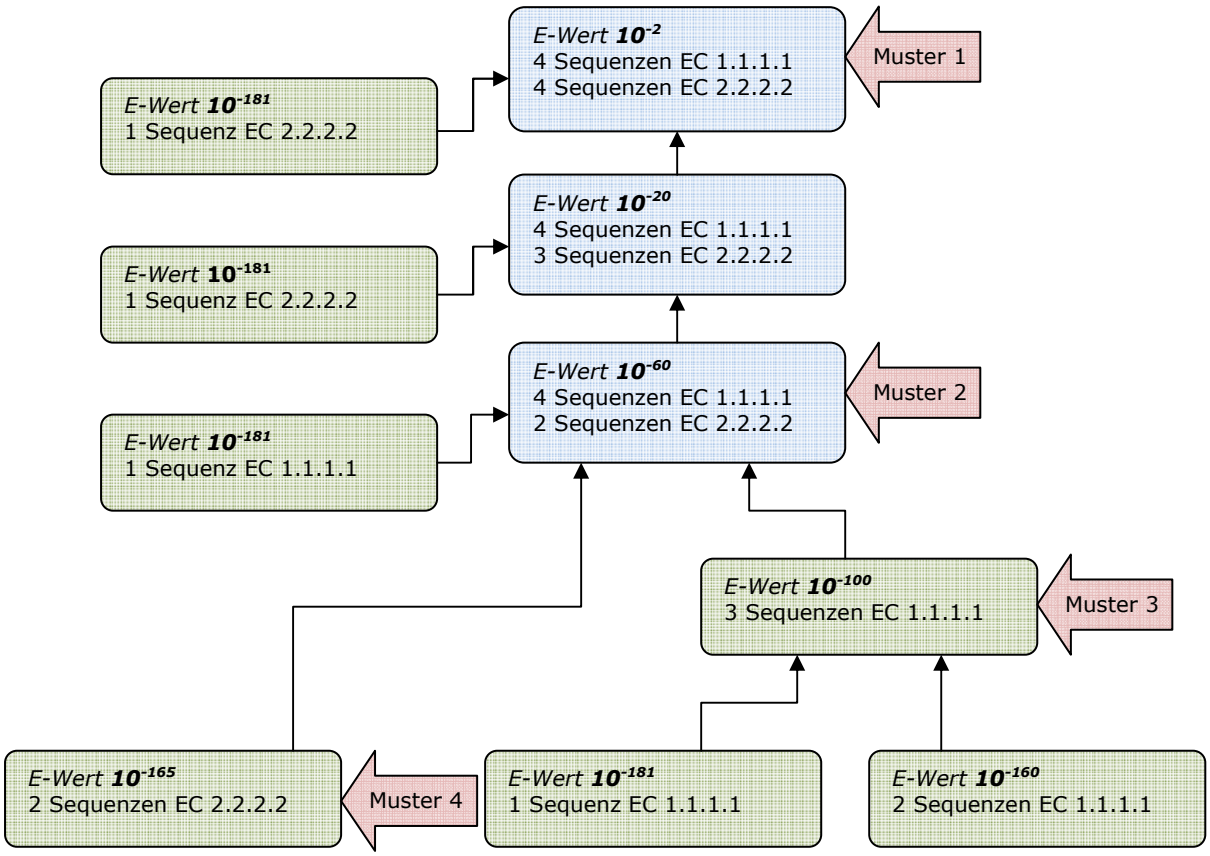


Abbildung 2-7: Ein beispielhafter Clusterbaum der Datenbank *tee*. An verschiedenen Stellen werden aus den Sequenzen der Cluster Muster erstellt.

Nach der Clusteranalyse sind Clusterbäume entstanden, dessen Cluster homologe Sequenzen enthalten. Es können nicht aus jedem Cluster Sequenzen für die Erstellung von Mustern entnommen werden, denn es würden zu viele Muster entstehen. Aus diesem Grund wurde bestimmt, dass aus den Sequenzen aus folgenden Clustern Muster generiert werden:

- Bei E-Wert 10^{-2}.
- Wenn sich die Anzahl der EC-Nummern im Cluster ändert, wird aus den Sequenzen dieses Clusters ein Muster erstellt, sowie aus allen Clustern, deren Sequenzanzahl größer eins ist und zu dem Cluster führen, indem sich die Anzahl der EC-Nummern geändert hat.

In Abbildung 2-7 ist ein Clusterbaum beispielhaft abgebildet. Cluster mit mehr als einer EC-Nummer sind blau, Cluster mit einer EC-Nummer sind grün unterlegt. Die Spitze des Clusterbaums bildet das Cluster bei E-Wert 10^{-2}. An den mit Pfeilen markierten Stellen wird aus den Sequenzen der Cluster ein Muster generiert. Nach Konvention, wird bei E-Wert 10^{-2} das Muster 1 erstellt (vgl. Abbildung 2-7). Muster 2 wird generiert, da sich die Anzahl der EC-Nummern im Cluster von einer EC-Nummer auf zwei EC-Nummern geändert hat. Muster 3 und Muster 4 werden erstellt, da aus diesen Clustern die Sequenzen stammen, aus dem Muster 2 erstellt wird und die Anzahl der Sequenzen in diesen Clustern größer eins ist.

Ist in einem Clusterbaum nur eine EC-Nummer vorhanden, wird aus den Sequenzen des Clusters bei E-Wert 10^{-2} und zusätzlich aus allen Clustern, die zum Cluster bei E-Wert 10^{-2} führen (und mindestens zwei Sequenzen enthalten), Sequenzmuster erstellt.

2.5.2 Übersicht über die Vorgehensweise bei der Ermittlung von Sequenzmustern

Derzeit existieren mehrere Algorithmen und Programme, die automatisiert Alignments von Proteinsequenzen durchführen. Sie unterscheiden sich hauptsächlich in ihren Methoden, Geschwindigkeiten und Qualitäten. Mit Hilfe von Sequenzalignments können Homologien zwischen ausgewählten Sequenzen ermittelt werden (vgl. Abschnitt 1.6).

Zu Beginn dieser Arbeit stand kein adäquates Programm zur Verfügung, das ein Sequenzmuster im PROSITE-Format anhand eines nicht-alignierten Sequenzdatensatzes erstellt. Das einzige Programm, das ein Sequenzmuster im PROSITE-Format ausgeben kann, ist PRATT [24]. Allerdings gibt PRATT nicht ein Muster, sondern mehrere Muster unterschiedlicher Qualität aus. Zudem kann PRATT sehr lange Sequenzen nicht verarbeiten bzw. es gibt eine maximale Sequenzzahl, die in einem Durchgang enthalten sein darf.

Aufgrund dieser Tatsache wurde zur Ermittlung von Sequenzmustern eine Strategie entwickelt, so dass konservierte Aminosäuren von mehreren nicht-alignierten Sequenzen identifiziert werden können. Die folgende Abbildung 2-8 gibt Aufschluss über einzelne Schritte dieses Verfahrens.

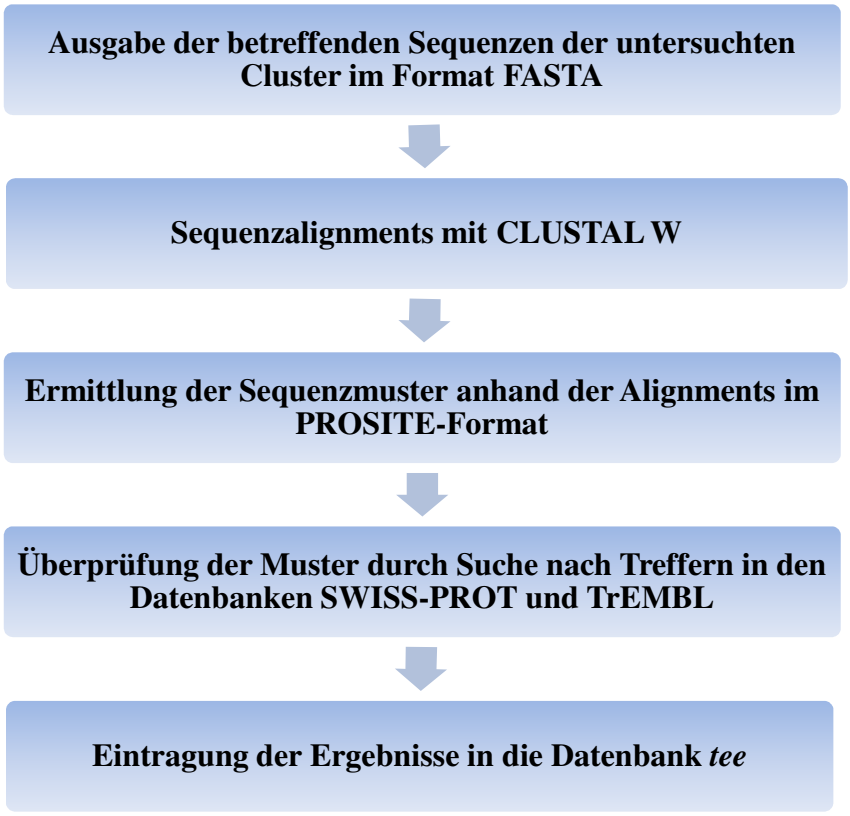


Abbildung 2-8: Übersicht über den Verlauf der Mustererstellung.

Im ersten Schritt werden die Cluster identifiziert, aus deren Sequenzen Muster erstellt werden sollen (vgl. Abschnitt 2.5.1 und Diagramm 2-7). Diese Sequenzen werden im Dateiformat FASTA gespeichert und im zweiten Schritt dem Programm CLUSTAL W übergeben, das multiple Alignments durchführt. Die konservierten Aminosäuren, die CLUSTAL W im Alignment markiert, werden genutzt, um Muster im PROSITE-Format zu generieren. Mit der Suche nach Treffern dieser Muster in den Datenbanken SWISS-PROT und TrEMBL wird die Qualität der Sequenzmuster überprüft. Die Ergebnisse dieser Überprüfung, sowie weitere Werte wie z.B. E-Wert, Sequenzanzahl und Clusterbaumnummer werden in den Tabellen der Datenbank *tee* gespeichert. Die folgenden Abschnitte geben eine detailierte Übersicht über Einzelheiten des Verfahrens.

2.5.2.1 Das FASTA-Format

Die Sequenzen der zu untersuchenden Cluster werden extrahiert und im Dateiformat FASTA gespeichert. Diesem Standard entsprechend werden die Sequenzen in einer Datei folgendermaßen eingetragen:

```
>ACJDNJ|3.4.4.3|5_40|3
VPDEAHQTFLLPCERCKYTWGGNTVHAHWLLRKDGY
>SDFFIF|3.6.7.1|2_39
TFLLPCKYGNHALRKGYAFLCERCKYTVVPDEAHQTFL
>SFKGDR|3.6.7.1|1_37
TLCKGNARKGFERKTVEALRGYFLERKLRGYALCRCY
```

Abbildung 2-9: Das FASTA-Format. Das Beispiel enthält drei Sequenzen.

Jede Aminosäuresequenz hat eine beschreibende Zeile, die mit dem „>“-Zeichen beginnt. Nach diesem Zeichen können beliebige Informationen folgen. Es empfiehlt sich aber, diesen Teil zur späteren Identifizierung der Sequenz sinnvoll zu nutzen. Das Beispiel zeigt, welcher Inhalt in dieser Dissertation als sinnvoll beurteilt wurde: Es befinden sich drei Informationen in dieser Zeile, die jeweils von einem „|“-Zeichen getrennt werden. Im ersten Teil ist die SWISS-PROT Accession-Nummer zu sehen (Erklärung zu SWISS-PROT siehe Abschnitt 2.11.1), gefolgt von der vierstelligen EC-Nummer, die dieser Sequenz zugeordnet wurde. Da Abschnitte (Domänen) von Enzymen untersucht werden, ist zu jedem Sequenzeintrag der untersuchte Bereich der Sequenz angegeben. Diese Information ist im letzten Teil der Zeile enthalten. Diese Datenstruktur gilt für alle Einträge dieser Datei. In der ersten beschreibenden Zeile ist eine zusätzliche Information verfügbar. Die Zahl an dieser Stelle beschreibt die Anzahl der vorhandenen Sequenzen in dieser Datei.

Im obigen Beispiel hat die erste Sequenz die Accession-Nummer ACJDNJ. Sie trägt die EC-Nummer 3.4.4.3. Der Sequenzabschnitt von Aminosäure fünf bis Aminosäure 40 wurde in die FASTA-Datei eingetragen. Im dargestellten Beispiel sind insgesamt drei Einträge zu sehen. Die Aminosäuresequenz befindet sich jeweils als Ein-Buchstaben-Code unterhalb der beschreibenden Zeile.

2.5.2.2 Globale Sequenzalignments mit CLUSTAL W

Die meisten Programme erstellen progressive multiple Alignments nach der Methode von Feng und Doolittle [9]. Die sog. dynamische Programmierung [131] erlaubt die Erstellung eines optimalen mathematischen Alignments (bei zwei Sequenzen) in akzeptabler Zeit. Bei mehreren längeren Sequenzen sind für diese Methode Computerressourcen nötig, die in der Praxis nicht vorhanden sind.

Das Computerprogramm CLUSTAL W wurde verwendet, um Sequenzalignments durchzuführen. Es wurde am *European Molecular Biology Laboratory* von Thompson *et al.* entwickelt und 1994 veröffentlicht [101]. Es lässt sich auf verschiedene Computer und Betriebssysteme installieren.

CLUSTAL W löst das Ressourcenproblem und erstellt Alignments in kurzer Zeit. Es hat sich als adäquates Mittel für die Zuordnung von unbekannten Sequenzen zu Proteinfamilien und als Hilfe bei der Identifizierung von Verwandtschaftsverhältnissen unterschiedlicher Sequenzen bewährt.

Die Rechenschritte, die CLUSTAL W während eines Alignments durchführt, werden in den folgenden Abschnitten erläutert. Die Angaben wurden der Veröffentlichung aus dem Jahr 1994 entnommen [101]. Die Berechnungen gliedern sich in drei Schritte. Zunächst wird eine Distanzmatrix berechnet. Im zweiten Schritt wird der *guide tree* erstellt. Schließlich wird ein progressives Alignment durchgeführt.

1. Schritt: Berechnung der Distanzmatrix:

Die Distanzmatrix wird zu Anfang des Durchlaufs mittels paarweisen Sequenzalignments berechnet. Dieses Verfahren ermöglicht es, eine große Sequenzanzahl ressourcenschonend miteinander zu vergleichen. Es werden *scores* berechnet, bei deren Erstellung *gap penalties* (Strafen für die Einfügung von Lücken im Alignment) abgezogen werden. Eine zweite, langsamere Methode zur Berechnung kann ausgewählt werden. Hier wird zwischen der Einfügung eines *gaps* und der Verlängerung eines *gaps* unterschieden. Zusätzlich wird eine Aminosäure *weight matrix* genutzt. Die *scores* werden dadurch berechnet, indem die Anzahl der identischen Aminosäuren im besten Alignment mit der Gesamtzahl der Aminosäuren dividiert wird. Beide *scores* geben den prozentualen Anteil der Identitäten an. Die Distanzen errechnet sich aus dem *score* geteilt durch einhundert subtrahiert von eins.

2. Schritt: Erstellung des *guide tree*

Mit Hilfe der zuvor ermittelten Distanzmatrix wird der *guide tree* erstellt. Das ist ein Stammbaum, der nach dem Verfahren der *Neighbour-Joining-Methode* [134] konstruiert wird. Im *guide tree* sind Längen der Abzweigungen proportional zu den geschätzten Abweichungen entlang jeder Abzweigung. Die Wurzel des Baums wird nach der *mid-point* Methode [102] an der Position gesetzt, wo das arithmetische Mittel der Längen der Abzweigungen auf jeder Seite des Baums gleich ist. Diese Bäume werden auch dazu genutzt, um eine Gewichtung für jede Sequenz zu erhalten [102]. Die Gewichtung resultiert aus der Distanz von der Wurzel des Baums, daher haben Sequenzen, die einen gemeinsamen Zweig des Baums haben auch eine gemeinsame Gewichtung.

3. Schritt: Progressives Alignment

In diesem letzten Schritt des Alignments werden nacheinander Sequenzen paarweise nach dem Schema des *guide tree* verglichen. Dabei wird mit den Sequenzen begonnen, die sich an den Spitzen des Baums befinden, bis die Wurzel des Baums erreicht wird. Für diese Alignments wird ein dynamischer Algorithmus verwendet, der eine Aminosäure *weight matrix*, und *penalties* für die Einführung und Verlängerung von *gaps*, berücksichtigt. Bei jedem Schritt werden zwei existierende Alignments oder Sequenzen aligniert. Wurden in einem vorherigen Schritt *gaps* eingeführt, bleiben diese bei allen folgenden Schritten bestehen. Falls im Verlauf des Alignments neue *gaps* eingeführt werden, kommen zusätzliche *gap penalties* zustande, selbst dann, wenn diese neuen *gaps* in Bereiche zuvor eingeführter *gaps* liegen. Um den *score* eines Vergleichs von einer Position eines Alignments und einer Position einer Sequenz zu bestimmen, wird der Durchschnitt aller paarweisen *weight matrix scores* von Positionen beider Elemente berechnet. Der *score* ist gleich Null, wenn ein *gap* mit einer Aminosäure verglichen wird.

Das progressive Alignment in diesem dritten Schritt des Verfahrens ist der wichtigste Schritt auf dem Weg zu einem korrekten, multiplen Alignment. Aus diesem Grund wurde versucht, diesen Schritt zu beschleunigen und die Genauigkeit des Vergleichs zu verbessern. Einfluss auf das Ergebnis haben während des progressiven Alignments:

- *Gap penalties:*

  Die Einführung von zwei unterschiedlichen *penalties*, dem GOP (*gap opening penalty*), wenn eine Lücke in das Alignment eingeführt wird und dem GEP (*gap extension penalty),* wenn eine bestehende Lücke erweitert wird, helfen, das multiple Alignment zu verbessern.

- *Weight matrix:*

  Die Wahl der *weight matrix* in Verbindung mit GOPs und GEPs beeinflusst die Genauigkeit des Alignments. In CLUSTALW stehen mehrere Matrizen zur Auswahl.

- Ähnlichkeit der Sequenzen:

  Die prozentuale Angabe der identischen Aminosäuren zweier Sequenzen wird dazu verwendet, die Werte der GOPs zu verbessern, um ein besseres Alignment zu erhalten, wenn sehr ähnliche, bzw. sehr verschiedene Sequenzen im Alignment verglichen werden.

- Länge der Sequenzen:

  Die erzielten *scores* werden umso länger, je länger die zu vergleichenden Sequenzen sind. Die Werte der GOPs werden angepasst, wenn sehr lange oder sehr kurze Sequenzen verglichen werden. Ebenso werden die Werte der GEPs angepasst, wenn der Längenunterschied zwischen Sequenzen sehr groß ist.

- Abhängigkeit der *gap penalties* von der Position der Lücke:

  Bei den meisten Programmen, die dynamische Algorithmen nutzen, sind die Werte für *gap penalties* unabhängig von dem Ort, an dem eine Lücke eingeführt wird. In CLUSTAL W werden die Werte für *gap penalties* angepasst, je nachdem wie wahrscheinlich es ist, dass an einer bestimmten Stelle eine Lücke eingeführt wird. Zudem werden die Werte für *gap penalties* angepasst, wenn eine Lücke…

  (1) …in die Nähe (innerhalb 8 Positionen) einer bereits bestehenden Lücke eingeführt wird.
  (2) …in einem hydrophilen Bereich eingeführt wird.
  (3) …bestimmte Aminosäuren betrifft.

Die folgende Abbildung zeigt ein Beispiel von sechs Sequenzen, die mit Hilfe von CLUSTAL W aligniert werden. Es soll das Prinzip einer CLUSTAL W Ausgabe erklären. Auf die Bedeutung der Symbole wird in Abschnitt 2.7 eingegangen.

```
1.Sequenz               EPLLKAATAGTIPLIPGISTVSELMLGMDYGLKEFKFFP
2.Sequenz               DELLKAACEGSIPLIPGISTVSELMLGMSYGLREFKFFP
3.Sequenz               ESLAKAGHNEIPGVATPGEIQLALEHGIDTLKLFP
4.Sequenz               KYIVAAANDSDVPPPPLGPAAITPGEMLALREEGYTHLKFFP
5.Sequenz               EDILEAGVDSEIPKLPGISTPSEIMMGYALGYRRFKLFP
6.Sequenz               DELLRFALDSEVPLLPGVASASEIMLAYRHGYRRFKLFP
```


```
1.Sequenz               EPLLKAATAGTIPLIPG---ISTVSELMLGMDYGLKEFKFFP
2.Sequenz               DELLKAACEGSIPLIPG---ISTVSELMLGMSYGLREFKFFP
3.Sequenz               ESLAKAGHN----EIPG---VATPGEIQLALEHGIDTLKLFP
4.Sequenz               KYIVAAANDSDVPPPPLGPAAITPGEMLALREEGYTHLKFFP
5.Sequenz               EDILEAGVDSEIPKLPG---ISTPSEIMMGYALGYRRFKLFP
6.Sequenz               DELLRFALDSEVPLLPG---VASASEIMLAYRHGYRRFKLFP
                        . :   .        *      : .*:      *   :*:**
```

Abbildung 2-10: CLUSTAL W erstellt aus sechs Sequenzen ein Alignment.

2.5.2.3 CLUSTAL W Version und Einstellungen

Während dieser Arbeit wurde die aktuelle CLUSTAL W Version 1.8.3 genutzt. Die verwendeten Einstellungen von CLUSTAL W entsprechen den Standardeinstellungen des Programms, da CLUSTAL W bei mehreren tausend Alignments nicht für jedes Alignment individuell eingestellt werden kann. Eine Übersicht der verwendeten Einstellungen findet sich im Anhang.

2.6 Löschung von Clustersequenzen

CLUSTAL W verarbeitet Datensätze, die mehr als 1000 oder mehr Sequenzen enthalten. Je mehr Sequenzen vorhanden sind, desto länger dauert das Alignment. Da Clusterknoten erwartet wurden, die deutlich mehr als 10000 Sequenzen enthalten, wäre ein Alignment mit dieser Sequenzanzahl zeitlich kaum zu realisieren. Ein Lösungsansatz zur Reduzierung der Sequenzanzahl in einem Alignment, ohne Informationen zu verlieren, bietet die Löschung von überlappenden Abschnitten einer Sequenz.

Auf diese Weise kann die Anzahl der Sequenzen in den untersuchten Clustern erheblich reduziert und Alignments können in einer höheren Geschwindigkeit durchgeführt werden, ohne Sequenzinformationen zu verlieren. Zusammenfassungen wurden nur durchgeführt, sofern ein Alignment erstellt wurde.

Beispiel:

Sind in einem Cluster die Sequenzabschnitte einer Sequenz

(1) Sequenzabschnitt 1: AS 1 bis AS 67
(2) Sequenzabschnitt 2: AS 65 bis AS 402
(3) Sequenzabschnitt 3: AS 106 bis AS 244

enthalten, wird Sequenzabschnitt 3 aus dem Alignment entfernt, da dieser Bereich komplett von Sequenzabschnitt 2 abgedeckt wird.

Zur Überprüfung dieser Methode wurden Alignments mit den Ausgangssequenzen und mit zusammengeführten Sequenzen durchgeführt. In allen untersuchten Fällen erzeugte CLUSTAL W identische Alignments.

Da CLUSTAL W aus Alignments mit mehr als 1000 Sequenzen in der Regel keine konservierten Aminosäuren erkennen kann, werden Alignments aus mehr als 1000 Sequenzen nicht durchgeführt.

2.7 Sequenzmuster im PROSITE-Format

Nach der Erstellung von Alignments, können aus diesen Alignments Sequenzmuster extrahiert werden. CLUSTAL W markiert identische Aminosäuren mit einem Stern (“*“). Diese Positionen werden als „hochkonserviert“ bezeichnet. Ähnliche Aminosäuren markiert CLUSTAL W mit einem Doppelpunkt(“:“). Diese Positionen werden als „konserviert“ bezeichnet. Aminosäuren, die CLUSTAL W mit einem Punkt (“.“) markiert, sind so schwach konserviert, dass entschieden wurde, diese Markierung bei der Erstellung von Mustern nicht zu berücksichtigen.

Die PROSITE-Datenbank verwendet zur Darstellung von Sequenzmustern ein etabliertes Standardformat, das auch in dieser Arbeit verwendet wird. Gemäß der Tabelle 2-1 findet folgende Syntax Verwendung:

Symbol	**Bedeutung**
Großbuchstabe	Aminosäure nach dem Ein-Buchstaben-Code
[Großbuchstaben]	Position mit mehreren möglichen Aminosäuren
X	eine beliebige Aminosäure
x(a)	eine Lücke mit a Aminosäuren
x(a,b)	eine Lücke mit a bis b Aminosäuren
Bindestrich "-"	bedeutungsloser Abstandshalter

Tabelle 2-1: Verwendete PROSITE-Syntax zur Darstellung von Sequenzmustern.

Beispiel:

In der folgenden Abbildung ist eine Ausgabe von CLUSTAL W dargestellt. Darunter befindet sich das aus diesem Alignment erzeugte Sequenzmuster. Das CLUSTAL W Alignment, bzw. die konservierten und markierten Aminosäuren im Alignment, dienen dabei als Grundlage. Die Syntax des PROSITE-Formats kann anhand der Tabelle 2-1 nachvollzogen werden.

Ausgabe von CLUSTAL W

```
Q10739|3.4.24.80|61_223|6       LSAAIAAMQRFYGLQVTGKADSDT--AMRRPRCGVPD
P53690|3.4.24.80|61_223         LSAAIAAMQKFYGLQVTGKADLAT--AMRRPRCGVPD
P33436|3.4.24.24|70_214         LKDTLKKMQKFFGLPQTG-LDQNT--TMRKPRCGNPD
P33434|3.4.24.24|70_214         LKDTLKKMQKFFGLPQTG-LDQNT--TMRKPRCGNPD
Q90611|3.4.24.24|67_211         LKDTLKKMQKFFGLPETGDLDQNTIETMKKPRCGNPD
P22757|3.4.24.12|131_292        ----ILDFQEHGGINQTGILDADTAELLSTPRCGVPD
                                    :  :*.. *:  **  *  *   :  **** **
```

Erzeugtes Muster

```
[IL]-x(2)-[FM]-Q-x(3)-G-[IL]-x(2)-T-G-x(1,2)-D-x(2)-T-x(1,3)-[LM]-x(2)-P-R-C-G-x-P-D
```

Abbildung 2-11: Auf Grundlage des CLUSTAL W Alignments wird ein Muster im PROSITE-Format erstellt.

Das abgebildete Muster hat die Länge 16, da Positionen und Lücken (x und x(0,∞)), an denen jede beliebige Aminosäure vorkommen kann, nicht gezählt werden.

Um die Qualität der Muster zu gewährleisten, werden generierte Muster, die eine Länge von acht Positionen unterschreiten, nicht in die Datenbank eingetragen.

2.8 Beurteilung des Musterbegriffs

In den folgenden Abschnitten wird auf die theoretische und praktische Bedeutung von Sequenzmustern eingegangen. Anhand von Beispielen wird erklärt, welche Definition in dieser Arbeit verwendet wird, welche Merkmale die biologische Qualität von Sequenzmuster beeinflussen und nach welchen Kriterien Treffer von Muster als Richtig-Positiv oder Falsch-Positiv beurteilt werden.

2.8.1 Definition des Musterbegriffs

Einige Algorithmen, die Sequenzmuster identifizieren, nutzten bei der Speicherung des Musters das PROSITE-Format oder Abwandlungen davon.

Arikawa *et al.* [113] nutzte das Prinzip der Indexierung als Instrument, Muster zu erkennen. Dabei wurde ein Drei-Buchstaben Alphabet für Aminosäuren konstruiert. Die Indexierung eines Alphabets $\sum$ ist die Funktion f: $|\sum| \rightarrow |\sum'|$ und führt dazu, dass Aminosäuren gruppiert werden. Musterpositionen, die eine einzelne Aminosäure beschreiben, existieren nicht. Damit ist dieser Ansatz sehr starr und nicht dazu geeignet, einzelne, flexible Positionen zu beschreiben.

Das entwickelte Mustermodell von Smith und Smith [37] ist flexibler. Es definiert einen Baum und die Aminosäuren bilden die Blätter dieses Baums. Es ist möglich, im Muster zunächst jede beliebige Aminosäure darzustellen (Symbol X) und Lücken sind im Muster erlaubt. Auf diese wird letztendlich ein PROSITE ähnliches Musterformat erstellt.

Sargot *et al.* [114] entwickelte einen Algorithmus, der in der Lage ist, Muster zu generieren. Es wurden Symbole eingeführt, die Gruppen von Aminosäuren entsprechen. Aufgrund des Suchverfahrens ist es nicht möglich, Muster zu erstellen, die Platzhalter enthalten.

Smith *et al.* [36] stellte ein Verfahren zur automatischen Entdeckung und Erstellung von Sequenzmuster vor. Das Musterformat ist vom Typ a_1-x(d_1)-a_2-x(d_2)-a_3, wobei a_1, a_2, a_3

Symbole des Alphabets und d_1, d_2 Ziffern repräsentieren. Das dargestellt Muster trifft Sequenzen mit a_1, a_2, a_3 und einer Lücke von d_1 und d_2 zwischen diesen Positionen. Ein ähnliches Verfahren wird in der PROSITE Syntax eingesetzt.

2.8.2 Definition als regulärer Ausdruck und Beispiele

Eine Zeichenkette, die der Beschreibung von Mengen bzw. Untermengen von Zeichenketten mit Hilfe syntaktischer Regeln dient, wird in der Informatik „regulärer Ausdruck“ genannt [108]. Reguläre Ausdrücke werden vor allem in der Softwareentwicklung verwendet. Für die in dieser Arbeit genutzte Programmiersprache Python ist das Modul *re* [109] vorhanden, mit dessen Hilfe reguläre Ausdrücke konstruiert werden können. Reguläre Ausdrücke dienen dazu, Texte zu filtern, indem diese in Form eines Musters mit den Wörtern des Textes verglichen werden. Es ist beispielsweise möglich, ein Text zu durchsuchen und alle Wörter zu finden, die mit „-ase“ enden, ohne ein spezielles Präfix bestimmen zu müssen. Auch können Wörter gefunden werden, die mit dem Buchstaben „A“ beginnen und mit dem Buchstaben „e“ enden, falls der reguläre Ausdruck entsprechend definiert ist.
In dieser Arbeit wird jedes erstellte Muster als regulärer Ausdruck genutzt, um in den Datenbanken SWISS-PROT und TrEMBL nach Treffern dieser Muster auf Elemente der Datenbanken zu suchen.

Beispiele:

Das Sequenzmuster

```
[IL]-x(2)-[FM]-Q-x(3)-G-[IL]-x(2)-T-G-x(1,2)-D-x(2)-T-x(1,3)-[LM]-x(2)-P-R-C-G-x-P-D
```

trifft die ersten beiden Sequenzen der Datenbank, die in Abbildung 2-12 dargestellt wird. Wie dem Diagramm zu entnehmen ist, unterscheiden sich die Aminosäuresequenzen, sowie die Längen der getroffenen Sequenzen. Da das Sequenzmuster als regulärer Ausdruck dient, werden beide Treffer identifiziert.

```
1.Sequenz          GDLLYKLEEFQACAGIPGTGGDVYTGGLMWPRCGHPDHFK
2.Sequenz          DELLDGWLKKIGGMQHILGLHHTGHTDYYTGHTLHHPRCGHPDFP
3.Sequenz          ESLAKAGLGGHNEIPGVATPGEIQLALEHGIDTLKLFPHKLIESGH
4.Sequenz          KYIVAAANDSDVPPPPGLGGLGPAAITPGEMLALREEGYTHLKFFP
5.Sequenz          EDILEAGVDSEIPKLPGISTPSEIMWHWPMGYALGYRRFKLFP
6.Sequenz          DELLRFALDSEVPLLPGVASASEIMLAYRHGYRGLGGRFKLFP
7.Sequenz          DELLRFALDSWHEVPLLPGVASASEIMLAYYGDRHGYRWPRFKLFP
8.Sequenz          DELLRFALDSEVPLLPGVASWHEIMLAYRHGYRKLPKLWPRFKLFP
```

Abbildung 2-12: Beispielhafte Sequenzdatenbank mit markierten Treffern unterschiedlicher Muster.

Je kürzer das Sequenzmuster ist, desto mehr Treffer sind zu erwarten. Das Sequenzmuster `G-L-G-G`, das aus vier Positionen besteht, trifft aus statistischer Sicht mehr Sequenzen einer Sequenzdatenbank, als ein längeres Sequenzmuster. Die Anzahl der Treffer ist auch von der Art der Aminosäuren abhängig. Sind sehr seltene Aminosäuren, wie Tryptophan (Ein-Buchstaben-Code W, vgl. Tabelle 2-2) im Muster vorhanden, kann ein relativ kurzes Muster dennoch sehr spezifisch sein. Das Muster `W-H-W-P` trifft in der Datenbank des Beispiels ausschließlich die fünfte Sequenz, obwohl das Muster genauso lang ist, wie das Muster `G-L-G-G`, das in dieser Datenbank drei Treffer erzielte. Neben der Länge und der Aminosäurezusammensetzung, spielen bei der Bestimmung der Spezifität auch variable Lücken eine Rolle. Variable Lücken werden im Muster als x(Ziffer,∞) angegeben und beschreiben den Sequenzbereich, in dem die benachbarten Aminosäuren der Lücken liegen müssen. Relativ lange Lücken innerhalb eines Musters machen ein Muster sehr flexibel. Dadurch steigt die Wahrscheinlichkeit, dass ein Treffer gefunden wird, an. Ohne Lücken müssen die Aminosäuren `W-H-W-P` direkt nebeneinander liegen, damit das Muster `W-H-W-P` eine Sequenz trifft. Ist in diesem Muster eine variable Lücke vorhanden, wie z.B.im Muster `W-H-x(0,30)-W-P`, dürfen zwischen Null und dreißig beliebige Positionen zwischen den Aminosäuren `W + H` und `W + P` liegen, damit dieses Muster trifft. Dieses Muster trifft die Sequenzen sieben und acht der Datenbank in Abbildung 2-12.

Ala (A) 8.56	Gln (Q) 3.90	Leu (L) 9.84	Ser (S) 6.81
Arg (R) 5.54	Glu (E) 6.06	Lys (K) 5.22	Thr (T) 5.59
Asn (N) 4.19	Gly (G) 7.06	Met (M) 2.41	Trp (W) 1.33
Asp (D) 5.27	His (H) 2.21	Phe (F) 4.05	Tyr (Y) 3.03
Cys (C) 1.35	Ile (I) 5.93	Pro (P) 4.82	Val (V) 6.65

Ala (A) 8.07	Gln (Q) 3.96	Leu (L) 9.67	Ser (S) 6.71
Arg (R) 5.48	Glu (E) 6.72	Lys (K) 5.89	Thr (T) 5.36
Asn (N) 4.06	Gly (G) 7.01	Met (M) 2.41	Trp (W) 1.10
Asp (D) 5.39	His (H) 2.28	Phe (F) 3.89	Tyr (Y) 2.95
Cys (C) 1.44	Ile (I) 5.90	Pro (P) 4.79	Val (V) 6.80

Tabelle 2-2: Prozentuale Anteile der Aminosäuren, die in den Datenbanken SWISS-PROT (oben) und TrEMBL (unten) enthalten sind [92, 135]. Die dargestellten Daten basieren auf Statistiken, die mit den Veröffentlichungen der Datenbanken zur Verfügung gestellt wurden. Für jede Aminosäure werden der Drei-Buchstaben-Code und der Ein-Buchstaben-Code, sowie der prozentuale Anteil der Aminosäuren der Datenbanken dargestellt.

2.9 Beurteilung von Richtig-Positiven und Falsch-Positiven Treffern

Aus der Datenbank *tee* werden alle Muster entfernt, die kürzer als acht Positionen sind. Durch diese Maßnahme soll die Qualität der Muster gewährleistet werden, denn es wird angenommen, dass zu kurze Muster zu viele Falsch-Positive Treffer erzielen (vgl. Abschnitt 2.8.2).

Die Qualität aller übrigen Muster der Datenbank *tee* wird eingeschätzt, indem überprüft wird, in welchem Maße Muster Treffer erzielen, wenn mit diesen in den Datenbanken SWISS-PROT und TrEMBL gesucht wird. Bei dieser Überprüfung wird zwischen den Datenbanken, aus denen die Sequenzen stammen, unterschieden. Stammen die Sequenzen eines Musters aus der Datenbank SWISS-PROT, wird mit dem resultierenden Muster in der Datenbank SWISS-PROT nach Treffern gesucht. Stammen die Sequenzen eines Musters aus der Datenbank TrEMBL, wird mit dem resultierenden Muster in der Datenbank TrEMBL nach Treffern gesucht. Stammen die Sequenzen eines Musters aus beiden Datenbanken, wird in beiden Datenbanken nach Treffern dieses Musters gesucht. In der Datenbank *tee* (vgl. Beispiele im Teil Ergebnisse) wird die Zugehörigkeit der Sequenzen mit SPROT für die Datenbank SWISS-PROT, TREMBL für die Datenbank TrEMBL und TRROT für Sequenzen beider Datenbanken angegeben. Die Beurteilung, ob ein Treffer Richtig-Positiv oder Falsch-Positiv ist, wird anhand der EC-Nummer vorgenommen. Ist keine EC-Nummer in der beschreibenden Zeile (vgl. Abschnitt 2.5.2.1) der Treffersequenz vorhanden, gilt der Treffer als Falsch-Positiv. Einem Muster wird die EC-Nummer zugeordnet, die im Cluster vorhanden ist, aus deren Sequenzen das Muster erstellt wurde. Sind in einem Cluster mehrere EC-Nummern vorhanden (vgl. Abbildung 2-7), erhält das Muster alle EC-Nummern dieses Clusters. Sequenzen können mehrere EC-Nummern besitzen. Mindestens eine EC-Nummer ist vollständig, zusätzliche EC-Nummern können *wildcards* (“-“) enthalten (vgl. anschließende Übersicht). Sequenzen mit ausschließlich unvollständigen EC-Nummern wurden nicht aus den Datenbanken SWISS-PROT und TrEMBL übernommen. In der folgenden Übersicht wird die Beurteilung von Treffern bei unterschiedlichen EC-Kombinationen dargestellt.

EC-Nummer des Musters:	**EC 1.2.3.4**		**Beurteilung des Treffers**
EC-Nummer des Treffers:	EC 1.2.3.4	→	Richtig-Positiv
EC-Nummer des Treffers:	EC 1.2.3.-	→	weder positiv noch negativ
EC-Nummer des Treffers:	EC 1.2.-.-	→	weder positiv noch negativ
EC-Nummer des Treffers:	EC 1.-.-.-	→	weder positiv noch negativ
EC-Nummern des Treffers:	EC 1.2.3.4 + EC 3.3.3.3	→	Richtig-Positiv
EC-Nummern des Treffers:	EC 1.2.3.- + EC 3.3.3.3	→	weder positiv noch negativ
EC-Nummer des Treffers:	EC 3.3.3.3	→	Falsch-Positiv

EC-Nummer des Musters:	**EC 1.2.3.4 + EC 2.2.2.-**		**Beurteilung**
EC-Nummer des Treffers:	EC 1.2.3.4	→	Richtig-Positiv
EC-Nummern des Treffers:	EC 1.1.1.1 + EC 2.2.2.2	→	Richtig-Positiv
EC-Nummern des Treffers:	EC 1.1.1.1 + EC 2.2.2.-	→	Richtig-Positiv
EC-Nummern des Treffers:	EC 1.1.1.1 + EC 2.2.-.-	→	weder positiv noch negativ

2.10 Der c-MCS Algorithmus und die Berechnung von R-Matrizen

Der c-MCS Algorithmus ist eine Variante einer Vielzahl von existierenden Algorithmen [88, 164-167] zur Untersuchung von größten gemeinsamen Teilstrukturen von Molekülen. Der in dieser Dissertation verwendete Algorithmus wurde von Markus Leber entwickelt und veröffentlicht [86]. Die folgenden Abschnitte, die den c-MCS Algorithmus und die Erstellung von Reaktionsmatrizen beschreiben, sind eine Zusammenfassung aus seiner Arbeit.

Der c-MCS Algorithmus ist eine Kombination des c-MCS und des McGregor Algorithmus [84]. Bei einem Vergleich von zwei Molekülen, wird zunächst mit dem c-MCS Algorithmus versucht, eine größte gemeinsame Teilstruktur zu errechnen. Ist dieser Schritt nicht erfolgreich, wird der Algorithmus nach McGregor eingesetzt. Dabei wird die Suche nach der größten gemeinsamen Teilstruktur nach c-MCS nicht abgebrochen, vielmehr wird die Suche nach c-MCS fortgesetzt und die Bindungen, die an die bereits gefundene größte gemeinsame Teilstruktur grenzen, mit dem Algorithmus nach McGregor berechnet. Die detailierte Vorgehensweise des c-MCS Algorithmus ist in Abbildung 2-13 dargestellt.

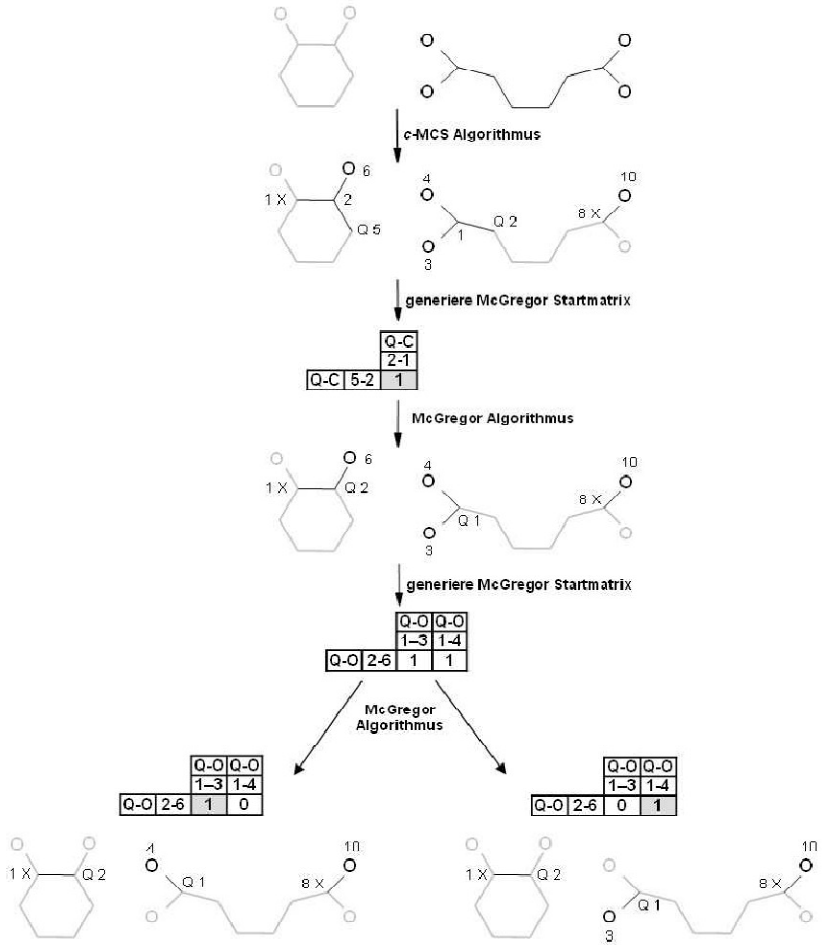


Abbildung 2-13: Funktionsprinzip des c-MCS Algorithmus‘, nach Leber [86].

Im ersten Schritt wird mit dem c-MCS Algorithmus die größte gemeinsame Teilstruktur bestimmt. Dann werden Bindungen, die noch nicht berücksichtigt wurden und an die größte gemeinsame Teilstruktur grenzen, an den McGregor Algorithmus übergeben. Damit die bereits erzielte Atomzuordnung eindeutig bleibt, wird die bisher erreichte Atomzuordnung konserviert, indem die Atome markiert werden. Im folgenden Schritt erweitert der McGregor Algorithmus solange die größte gemeinsame Teilstruktur, bis sich die Teilstruktur nicht mehr erweitern lässt. Am Ende der Berechnung werden zwei c-MCS ausgegeben. Da der McGregor Algorithmus ebenfalls die anderen Teilstrukturen ausbaut, werden am Ende der Berechnung acht c-MCS ausgegeben. Ist bei der Suche nach der größten gemeinsamen Teilstruktur der Zwischenschritt der Erweiterung der c-MCS durch McGregor nicht so einfach möglich wie in Abbildung 2-13 dargestellt, werden alle Möglichkeiten getestet, um die Substruktur zu erweitern. Auf diese Weise wird gewährleistet, dass das entwickelte Verfahren genau, schnell und flexibel ist [86].

2.10.1 Atom-Zuordnung biochemischer Reaktionen nach dem c-MCS Algorithmus

Der c-MCS Algorithmus (vgl. Abschnitt 2-10) wurde in dieser Arbeit dazu genutzt, die Moleküle, die in den von Enzymen katalysierten chemischen Reaktionen enthalten sind, miteinander zu vergleichen. Bei jedem Vergleich wird mindestens eine größte gemeinsame Teilstruktur ermittelt. Entstehen bei einem Molekülvergleich mehr als eine c-MCS, werden die resultierenden c-MCS untereinander verglichen, damit ein Set an c-MCS erhalten wird, bei dem alle Atome der verglichenen Moleküle zugeordnet werden können. Die Anzahl der erhaltenen c-MCS nimmt zu, je mehr Moleküle miteinander verglichen werden. Um alle Moleküle einer sehr umfangreichen chemischen Reaktion in einer angemessenen Zeit miteinander vergleichen zu können, wurden Verfahren entwickelt, um die Rechenzeit zu reduzieren. Zum einen trägt ein Rankingsystem dazu bei, c-MCS zu beurteilen, um eine korrekte Atomzuordnung zu erhalten. Auch heuristische Verfahren sind dabei behilflich, den Molekülvergleich zu beschleunigen, indem redundante c-MCS und lokale Symmetrien innerhalb von Molekülen erkannt und die betroffenen c-MCS entfernt werden. Zusätzlich zu diesen Verfahren, wird die Genauigkeit und Geschwindigkeit der Atomzuordnung dadurch gewährleistet, indem Spezialfälle, wie z.B. Co-Faktoren oder Moleküle, die komplizierte oder aromatische Ringe enthalten, gesondert behandelt werden. Um die Zuordnung von Molekülen und Atomen trennen zu können, wurde bei der c-MCS Berechnung zusätzlich eine Variante des Bron-Kerbosch Algorithmus [87] genutzt.

2.10.2 Beispiel für die Berechnung der größten gemeinsamen Teilstruktur

Im vorherigen Abschnitt wurde die theoretische Vorgehensweise des c-MCS Algorithmus bei der Suche nach einer größten gemeinsamen Teilstruktur beim Vergleich zwischen Molekülen beschrieben. Anhand eines Beispiels werden die Schritte nun erläutert. Das Enzym Pyruvatdecarboxylase katalysiert die chemische Reaktion der Decarboxylierung des Moleküls Pyruvat. Es entstehen die Produktmoleküle Kohlendioxid und Acetaldehyd. Der paarweise Vergleich der Kombinationen Pyruvat und Kohlendioxid, sowie Pyruvat und Acetaldehyd wird in Abbildung 2-14 dargestellt.

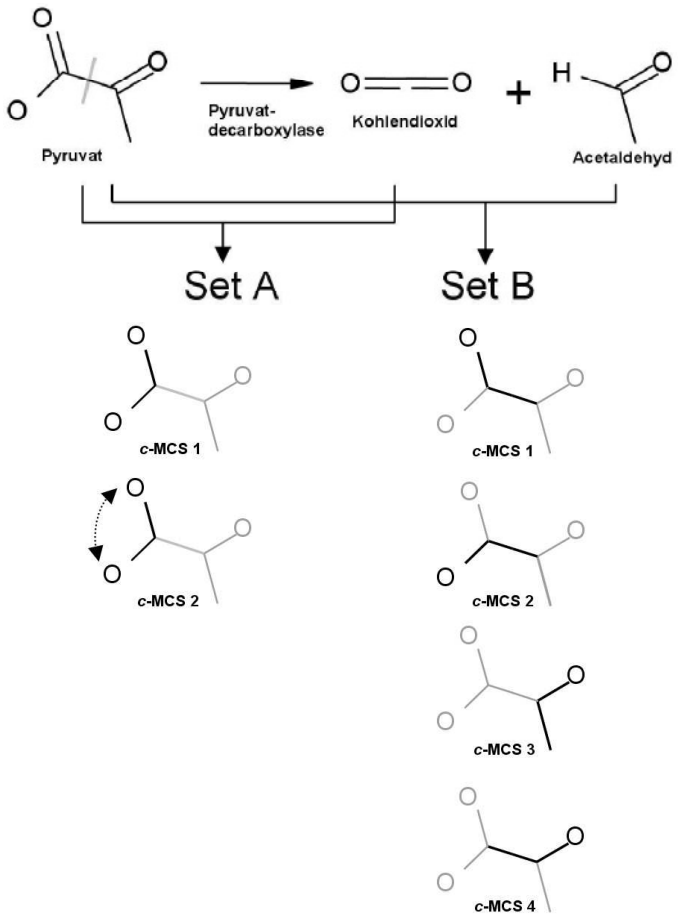


Abbildung 2-14: c-MCS Suche anhand der vom Enzym Pyruvatdecarboxylase katalysierten Reaktion, nach Leber [86].

Im Verlauf der c-MCS Suche werden im ersten Vergleich Pyruvat und Kohlendioxid miteinander verglichen, wobei zwei c-MCS generiert werden. Das Resultat des zweiten Vergleichs, Pyruvat und Kohlendioxid, besteht aufgrund der geringen Atomzahl des Kohlendioxids aus vier c-MCS. Nach der Generierung aller c-MCS, werden, um eine korrekte Atomzuordnung zu erhalten, alle c-MCS aus Set 1 mit allen c-MCS aus Set 2 verglichen. Daraus entstehen acht Kombinationen, wobei nur zwei Kombinationen zu vollständigen Atomzuordnungen führen.

2.10.3 Berechnung der Reaktionsmatrix (R-Matrix)

Reaktionsmatrizen (R-Matrizen) sind mathematische Operatoren, die Elektronentransfermuster von chemischen Reaktionen speichern [86]. In ihnen sind die Informationen gespeichert, welche Atome im Molekül vorhanden sind und welche Bindungen zwischen den Atomen gebildet oder gespalten werden. Diese Matrizen können dazu genutzt werden, chemische

Reaktionen, z.B. Reaktionen, die von Enzymen katalysiert werden, miteinander zu vergleichen [86].

Der beschriebene c-MCS Algorithmus bildet die Grundlage, automatisiert R-Matrizen zu erstellen. Nachdem eine Atomzuordnung bestimmt wurde, kann eine Reaktionsmatrix nach der Gleichung

Reaktionsmatrix = Produktmatrix - Eduktmatrix

erstellt werden. Anhand des Beispiels der Pyruvatdecarboxylase wird die Konstruktion der R-Matrix erklärt. Im ersten Schritt werden mit Hilfe des c-MCS Algorithmus` die Atome der zu vergleichenden Moleküle zugeordnet, eindeutig nummeriert und aus der Edukt- und Produktmatrix die Reaktionsmatrix erstellt.

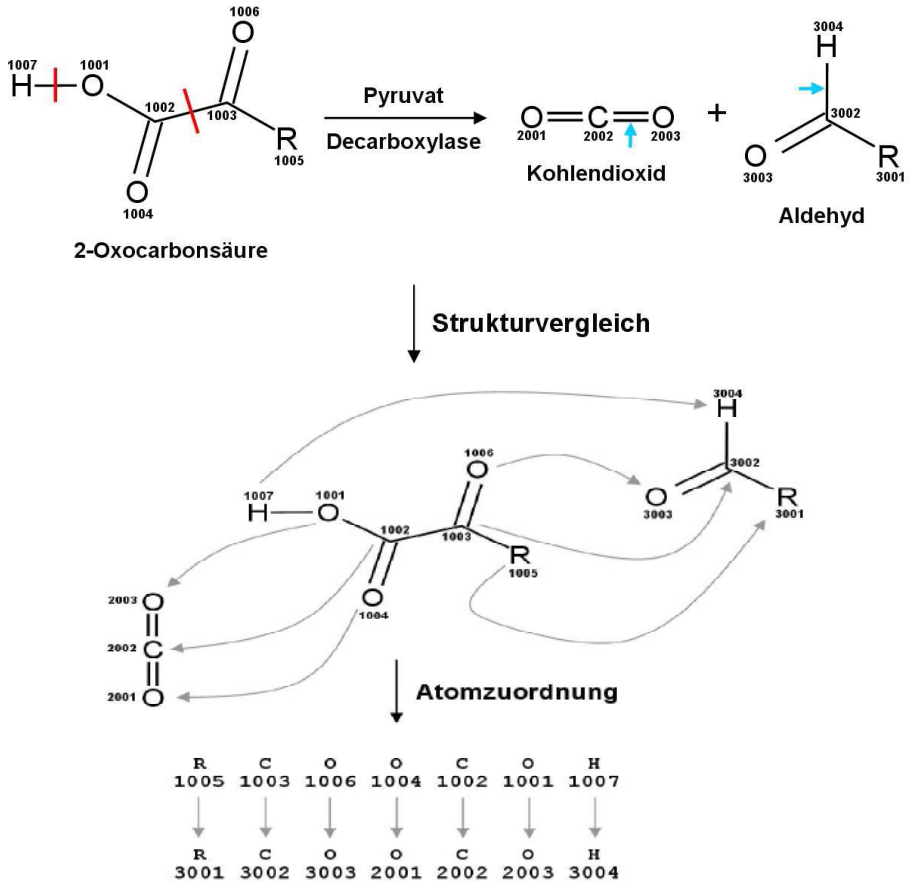


Abbildung 2-15: Die Erstellung der Reaktionsmatrix wird anhand der von dem Enzym Pyruvatdecarboxylase katalysierten Reaktion erklärt, nach Leber [86].

Die Matrizen werden nach dem in Abbildung 2-15 dargestellten Prinzip erstellt. Bindungen, die gebildet werden, erhalten positive Einträge. Bindungen, die gespalten werden, erhalten negative Werte. Da in einer Reaktion gespaltene Bindungen wieder eine Bindung eingehen, ist

die Summe aller Werte der Reaktionsmatrix gleich Null. Im Beispiel ist zu sehen, dass im Eduktmolekül zwei Bindungen gespalten werden, während zwei Bindungen im Produktmolekül entstehen. In der folgenden Darstellung sind die erstellten Matrizen abgebildet. Zeilen und Spalten, die nur aus Nullen bestehen, werden aus den Matrizen entfernt. Die Reaktionsmatrix enthält nur die an der Reaktion beteiligten Atome. Atome, die nicht an der Reaktion beteiligt sind, werden entfernt. In der ersten Spalte und Zeile sind die Atome des Eduktmoleküls dargestellt.

Produkt Matrix E

	R 3001	C 3002	O 3003	O 2001	C 2002	O 2003	H 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 1005	C 1003	O 1006	O 1004	C 1002	O 1001	H 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

	C 1003 C 3002	C 1002 C 2002	O 1001 O 2003	H 1007 H 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

2.10.4 Clusterung von Subsubklassen nach gleichen Reaktions-Strings (R-Strings)

Kanonisierte Reaktionsmatrizen lassen sich vereinfacht als Reaktions-Strings darstellen. Der Vorteil von R-Strings, die in einer einfachen Zeichenfolge gespeichert werden, ist die einfache Vergleichbarkeit der R-Strings untereinander. Dies kann dazu genutzt werden, R-Strings nach ihrem Elektronentransfermuster zu gruppieren.

Die R-Strings für 228 Subsubklassen wurden von Leber [86] untersucht. Für jede Subsubklasse wurde der häufigste R-String berechnet. 121 Subsubklassen besitzen ein spezifisches Elektronentransfermuster, die übrigen 107 Subsubklassen wurden nach identischen R-Strings geclustert. Auf diese Weise sind 26 Cluster unterschiedlicher Größe entstanden. Die Zusammensetzung der erhaltenen Cluster, sowie die Angaben über gespaltene und neue Bindungen sind in der folgenden Tabelle 2-3 zusammengefasst.

Gruppen-nummer	**Subsubklassen**		**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 2-3: Clusterung der EC-Subsubklassen nach identischen R-Matrizen, nach Leber [86].

2.10.5 Auswahl der EC-Kombinationen für die Analyse der katalysierten Reaktionen mit Hilfe des c-MCS Algorithmus‘

Die Auswahl der EC-Kombinationen für die Analyse der katalysierten Reaktionen mit Hilfe des c-MCS Algorithmus‘ bzw. der Vergleich der Reaktionen nach gleichen R-Strings, wird anhand des Clusterbaums erklärt, der in folgender Abbildung 2-16 dargestellt wird.

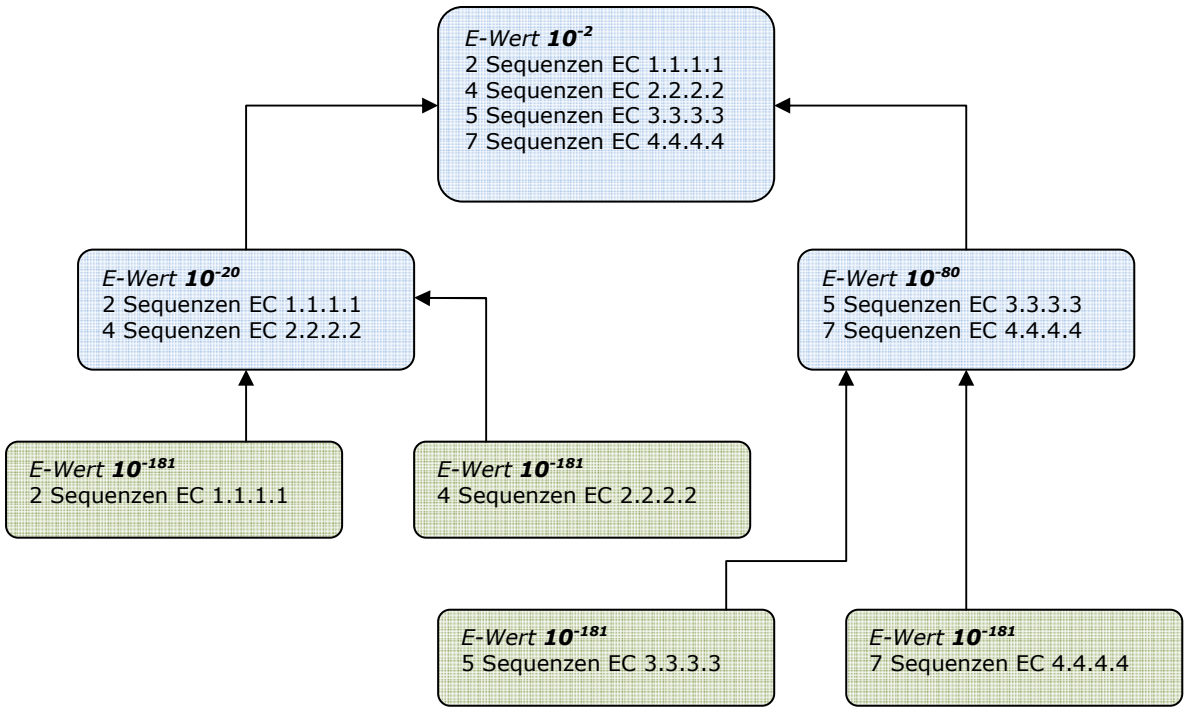


Abbildung 2-16: Bestimmung der EC-Kombinationen für die Analyse der größten gemeinsamen Teilstruktur der bei den Reaktionen beteiligten Moleküle, sowie für die Untersuchung der Zugehörigkeit zu Clustern basierend auf identischen R-Strings.

Bei verschiedenen E-Werten wird die EC-Zusammensetzung aller Cluster analysiert, die bei jeweiligem E-Wert mehr als eine EC-Nummer enthalten. Dabei wird die Reaktion jeder EC-Nummer mit jeder Reaktion aller anderen EC-Nummern dieses Clusters verglichen.

Anhand des Diagramms 2-16 entstehen folgende EC-Vergleiche:

Vergleiche bei E-Wert 10^{-2}:	EC 1.1.1.1 mit EC 2.2.2.2 EC 1.1.1.1 mit EC 3.3.3.3 EC 1.1.1.1 mit EC 4.4.4.4 EC 2.2.2.2 mit EC 3.3.3.3 EC 2.2.2.2 mit EC 4.4.4.4 EC 3.3.3.3 mit EC 4.4.4.4
Vergleich bei E-Wert 10^{-20}:	EC 1.1.1.1 mit EC 2.2.2.2
Vergleich bei E-Wert 10^{-80}:	EC 3.3.3.3 mit EC 4.4.4.4

Vergleiche von EC-Nummern, die sich ausschließlich in ihrer Seriennummer (vierte Stelle der EC-Nummer) unterscheiden, werden nicht durchgeführt, da sich die Reaktionen nur anhand der verwendeten Substrate unterscheiden.

2.10.6 Berechnung der größten gemeinsamen Teilstruktur

Die größte gemeinsame Teilstruktur von zwei Molekülen wird mit dem c-MCS Algorithmus berechnet (vgl. Abschnitt 2.10). Sind zwei Moleküle nicht identisch, wird der Anteil der größten gemeinsamen Teilstruktur in Prozent angegeben. Dieser Angabe liegt folgende Berechnung zu Grunde:

Anzahl Atome Molekül A: 10

Anzahl Atome Molekül B: 20

Anzahl gemeinsame Atome: 5

Im Vergleich von Molekül A mit Molekül B sind im Fall von Molekül A fünf von zehn Atomen identisch. Im Fall von Molekül B sind fünf von zwanzig Atomen identisch. Daraus ergeben sich die prozentualen Werte für die größte gemeinsame Teilstruktur von 50% für Molekül A und 25% für Molekül B.

In einer chemischen Reaktion sind typischerweise mehr als zwei Moleküle vorhanden. Dadurch entstehen mehrere Vergleiche, denn jedes Molekül, das in einer Reaktion vorhanden ist, wird

mit jedem Molekül aus der zu vergleichenden Reaktion verglichen, denn grundsätzlich katalysieren Enzyme auch die Rück-Reaktionen, so dass Produkte auch als Substrate fungieren können. In den Tabellen im Ergebnisteil wird für eine EC-Kombination aus allen Molekülvergleichen der jeweils größte prozentuale Wert eingetragen.

2.11 Verwendete Datenbanken

2.11.1 SWISS-PROT

Die Datenbank SWISS-PROT [135] (http://www.expasy.org/sprot), 1986 gegründet, enthält annotierte Proteinsequenzen im standardisierten Format [136]. Die frei zugängliche Datenbank enthält zu jeder Sequenz ausführliche Informationen über z.B. Sekundär- bzw. Quartärstruktur, Funktionen des Proteins, Ähnlichkeiten zu anderen Proteinen, Domänen und mögliche posttranslationale Modifikationen. Die Datenbank wird manuell annotiert und regelmäßig aktualisiert.

2.11.2 TrEMBL

Die Datenbank TrEMBL [92] (http://www.expasy.org/trembl) enthält computerannotierte Proteinsequenzen, die aus Translationen von allen codierenden Bereichen der DDBJ/EMBL/GenBank Nucleotid Datenbank gewonnen werden. Teilweise stammen Sequenzen auch aus speziellen Annotationsprogrammen, aus manueller Annotation und aus der Literatur. Die Qualität der Daten in TrEMBL hängt damit direkt von der Qualität der zur Verfügung gestellten Nucleotid Datenbanken und der verwendeten Algorithmen ab. Anders als die Datenbank SWISS-PROT liegt der Vorteil bei TrEMBL nicht in der Genauigkeit, sondern in der Geschwindigkeit der zur Verfügung gestellten Daten. TrEMBL wird in regelmäßigen Abständen aktualisiert.

2.11.3 PROSITE

PROSITE [137] (http://www.expasy.ch/prosite) ist eine öffentliche Datenbank, die Muster bzw. Profile von Proteinen enthält. Diese biologisch relevanten Muster sind die Grundlage für die Zuordnung unbekannter Proteine zu Proteinfamilien, deren Sequenz z.B. aus cDNA translatiert wurde. Zu den Einträgen in PROSITE sind kurze Dokumentationen enthalten; auf relevante Datenbanken, zum Beispiel PDB oder SWISS-PROT, wird verwiesen. Für weitere Untersuchungen sind mehrere Programme vorhanden, die es erlauben, eigene Sequenzen einzugeben, um sie einer Proteinfamilie zuordnen zu können [138]. Zusätzliche Informationen liefern Profile, die auf Grundlage von Sequenzalignments entstanden sind. Anders als Muster, decken Profile wesentlich längere Sequenzabschnitte ab und identifizieren auf diese Weise Proteindomänen oder Familien. PROSITE Release 20.30 vom 19. März 2008 enthält 1318 Sequenzmuster und 783 Profile.

2.11.4 PDB

Die Protein-Datenbank PDB [139] (http://www.rcsb.org/pdb) enthält dreidimensionale Strukturdaten von biologischen Makromolekülen, hauptsächlich Proteinen. Die Daten wurden größtenteils mit Hilfe von NMR oder Röntgenstrukturanalyse gewonnen. Die gespeicherten Daten lassen sich mit verschiedenen Programmen z.B. mit Pymol (vgl. Abschnitt 2.13.2) darstellen. Im April 2008 besteht die Datenbank aus 49974 Strukturen. PDB Daten werden zur Darstellung von Beispielen im dritten Teil dieser Arbeit verwendet.

2.11.5 BRENDA

BRENDA [140] (BRaunschweiger ENzyme DAtabase, http://www.brenda-enzymes.info/) ist eine relationale Datenbank, die molekulare und funktionale Informationen von Enzymen enthält. Die Daten der Enzyme stammen aus der Primärliteratur und werden ständig aktualisiert. Die Datenbank hat sich zu einem wichtigen Instrument in der Forschung entwickelt, da sich aus dieser Datenbank unter anderem Informationen über EC-Nummer, katalysierte Reaktion, Substrate/Produkte, Inhibitoren, Aktivatoren, Struktur und Stabilität abrufen lassen. Die in dieser Arbeit verwendeten Molfiles (Darstellungen der Moleküle in Dateien) wurden dieser Datenbank entnommen. Im aktuellen Release 2007.2 enthält BRENDA Daten zu 4757 EC-Nummern.

2.11.6 SCOP, CATH, Pfam und PRODOM

SCOP, CATH, Pfam und PRODOM sind öffentlich zugängliche Datenbanken, die Informationen zu biologischen Sequenzen liefern.

Die Datenbank SCOP [107] liefert detailierte Beschreibungen von strukturellen und evolutionären Beziehungen von allen Proteinen mit bekannter Struktur. Die Klassifizierung geschieht manuell durch visuelle Untersuchungen und Strukturvergleiche. Computerprogramme helfen dabei, Entscheidungen zu treffen. Die Ebenen der Klassifizierung bilden die Familie, Superfamilie und Faltung. Aufgrund der manuellen Klassifizierung kann die Einteilung subjektiv sein, dennoch bildet die SCOP Datenbank eine zuverlässige Klassifizierung, die häufig dazu genutzt wird, die Ergebnisse neuer Methoden zu überprüfen.

Die Datenbank Pfam [45] ist eine Sammlung von multiplen Alignments und *Hidden Markov Models* und liefert Informationen zur Domänenstruktur vieler Proteine. Pfam besteht aus den Teilen Pfam-A und Pfam-B. Während Pfam-A gut charakterisierte Domänen zusammenfasst, wird Pfam-B automatisch erstellt.

Die Datenbank CATH [81] ist eine Klassifizierung von Proteindomänen, die auf vier Ebenen vorgenommen wird: Klasse, Architektur, Topologie und Homologe Superfamilie. Die Klassifizierung der Klasse basiert auf Informationen von Sekundärstrukturen und wird für einen Großteil der Einträge automatisch vorgenommen. Die Einteilung der Architektur geschieht dagegen manuell. Die Architektur wird als grobe Orientierung der Sekundärstrukturen definiert. Die Stufe der Topologie wird aufgrund topologischer Verbindungen gebildet. Mittels Sequenz- und Strukturvergleiche wird die Klassifizierung der Homologen Superfamilien erreicht.

Die Datenbank PRODOM [58] enthält Informationen über Familien von Proteindomänen, die automatisch aus den Sequenzdatenbanken SWISS-PROT und TrEMBL gewonnen werden. Sequenzen werden geclustert und das Ergebnis ausgewertet. Auf diese Weise werden die evolutionären Beziehungen zwischen homologen Proteinen analysiert.

2.12 Erstellte Datenbanken

2.12.1 Die Datenbank *seq*

Diese Datenbank wurde auf Grundlage von SWISS-PROT und TrEMBL erstellt und enthält neben bereits vorhandenen enzymrelevanten Einträgen auch die komplette Clusterung der Sequenzen. Die Verbindung der Clusterdaten mit der Enzymdatenbank ist von großer Bedeutung, da sie die Basis für einen Großteil der Arbeit, sowie für die Erstellung der Datenbank *tee,* ist. Insgesamt wurden 47 Tabellen erstellt, die im relationalen Verhältnis zueinander stehen.

2.12.2 Die Datenbank *tee*

Tee ist eine Datenbank, die unter anderem aus den Tabellen *tree*, *edges* und *ecnumbers* besteht. Diese Datenbank enthält Informationen über die Beschaffenheit, Baumstruktur, enthaltene Sequenzen, Sequenzanzahl, E-Werte und EC-Nummern aller Clusterbäume. Auch alle nötigen Informationen um Clusterbäume grafisch darzustellen (Knotenkoordinaten, Clusterzusammensetzung und deren Verbindungen untereinander), befinden sich in diesen Tabellen. Eine Übersicht dieser und weiterer Tabellen, die während dieser Arbeit erstellt wurden, findet sich im Anhang.

2.13 Verwendete Programme und Programmiersprachen

Für die vorliegende Arbeit wurden Programme geschrieben, die spezielle Aufgaben erfüllen. Zum Beispiel sollte die Extraktion und Speicherung aller Sequenz- und Clusterdaten in der Datenbank automatisch ausgeführt werden. Als Programmiersprache wurde Python [141] verwendet. Neben eigenen Programmen wurden im Laufe dieser Arbeit auch bereits existierende Programme genutzt.

2.13.1 Das EMBOSS Paket und *patmatdb*

Das kostenlose EMBOSS-Paket [93] ist eine Sammlung von bioinformatischen Programmen, die sich dazu nutzen lassen, Sequenzalignments durchzuführen, Datenbanken zu durchsuchen, sowie Muster und Domänen von Proteinen zu analysieren. In der vorliegenden Arbeit wurde

das aus EMBOSS stammende Programm *patmatdb* genutzt, um mit Sequenzmustern nach Treffern dieses Musters in den Datenbanken SWISS-PROT und TrEMBL zu suchen.

2.13.2 PyMOL

PyMOL [142] wurde von DeLano Scientific LLC entwickelt und veröffentlicht. Das Programm nutzt PDB-Strukturdateien, um Moleküle, z.B. Proteine, dreidimensional darzustellen. PyMOL ist im Internet mittlerweile in der Version 1.1 frei verfügbar. Das Programm wurde bei der Analyse von Beispielen in Abschnitt 3 eingesetzt, um Proteinmoleküle abzubilden und um Musterpositionen innerhalb des Moleküls zu markieren.

2.13.3 yFiles und yED

yFiles [143] ist ein Programmpaket der Softwarefirma yWorks GmbH und gliedert sich in mehrere Einzelprogramme, die der Visualisierung unterschiedlichster Dateiformate dienen. Aus yFiles wurde das Programm yEd verwendet, um hierarchische Clusterbäume darzustellen. Das Programm wurde wegen seiner Schnelligkeit und Darstellungsmöglichkeit von Texten innerhalb der Clusterknoten ausgewählt. Die aktuelle Version 2.5.0.4 ist im Internet frei verfügbar.

2.14 Eigene Programme, Programmiersprache und Entwicklungsumgebung

Während der Dissertation wurden Computerprogramme entwickelt, die dazu dienten, Sequenzdaten automatisiert zu analysieren und zu verarbeiten. Die Ergebnisse wurden in verschiedenen Datenbanken gespeichert. Als Programmiersprache wurde *Python 2.5.1* verwendet, die objektorientierte, aspektorientierte und funktionale Programmierung unterstützt [103]. Zur Speicherung der Ergebnisse wurde die Open-Source-Software *MySQL,* ein relationales Datenbankverwaltungssystem, in der Version 5.0.18 eingesetzt [104]. Die Programme, die im Laufe dieser Arbeit entwickelt wurden, sind teilweise am Rechenzentrum der Universität zu Köln [105] und teilweise an einem Desktop PC (x86, 2,6 Gigahertz, 1024 Megabyte RAM) entstanden. Die Berechnungen der Datenbanken *tee* und *seq* wurden am Rechencluster der Biochemie der TU Braunschweig (9 Knoten mit je vier 2000 Megahertz 64 Bit AMD Prozessoren, je vier Gigabyte RAM) [106] durchgeführt.

Die entwickelten Programme und Datenbanken, sowie die genutzten Datenbanken SWISS-PROT und TrEMBL befinden sich auf der beiliegenden DVD. Eine Kurzkommentierung zur Funktionsweise und Anforderungen findet sich teilweise in den Dokumentationen der jeweiligen Programme zu Beginn des Quelltexts.

3 Ergebnisse

3.1 Die Clusterung der Sequenzen

3.1.1 Rechenzeitbedarf zur Erstellung der Clusterung

Die Clusterung der Sequenzen und die Erstellung der Datenbank *tee* wurden auf dem Computercluster in Braunschweig durchgeführt. Der Computercluster besitzt 9 Rechenknoten mit je vier CPUs des Herstellers AMD. Jeder Prozessor ist mit 2000 MHz getaktet. Ein Master-Rechenknoten steht für Berechnungen nicht zur Verfügung, so dass 32 CPUs von einem Anwender genutzt werden können.

Den größten Rechenbedarf bei der Erstellung des Datensatzes haben die lokalen all-vs-all BLAST Alignments. Nach Erhalt dieser Alignments, spielt beim Rechenbedarf die darauffolgende Clusterung der Sequenzen eine große Rolle. Hier ist weniger die Prozessorgeschwindigkeit, sondern ein großer Arbeitsspeicher für die Geschwindigkeit der Clusterung verantwortlich.

Abhängig von Faktoren wie die Datenbankgrößen von TrEMBL und SWISS-PROT, Zugänglichkeit zu Computerressourcen und Auslastung des Clusters durch Mitarbeiter, ist eine Clusterung mit den aktuellen Datenbanken von TrEMBL und SWISS-PROT auf einem oben beschriebenen Rechencluster in ca. 2 Wochen möglich.

3.1.2 Eigenschaften der Datenbank *tee*

Die Datenbank SWISS-PROT besteht aus 261513 Sequenzen, 75595 Sequenzen enthalten mindestens eine vollständige EC-Nummer und wurden in die Datenbank *tee* aufgenommen. Die Datenbank TrEMBL besteht aus 3987044 Sequenzen. Aus dieser Datenbank wurden 164328 Sequenzen extrahiert, die mindestens eine EC-Nummer enthalten. In der erstellten Datenbank *seq* befinden sich somit 239923 Sequenzen, entsprechend Accession-Nummern von SWISS-PROT und TrEMBL. Den extrahierten Sequenzen werden 2772 unterschiedliche EC-Nummern zugeordnet. Haben Sequenzen neben einer vollständigen EC-Nummer weitere EC-Nummern, die nicht vollständig sind, werden auch diese in die Datenbank aufgenommen. 86 EC-Nummern der Datenbank sind nicht vollständig.

Die Clusterergebnisse, das heißt die Zugehörigkeit der Domänen der jeweiligen Cluster bei E-Wert 10^{-2} bis E-Wert 10^{-181}, wurden in die Datenbank *tee*, bestehend u.a. aus den Tabellen

tree, ***e***dges und ***e***cnumbers, eingetragen. Ziel dieser Zusammenfassung war, die Beziehungen der Cluster untereinander zu verdeutlichen, so dass die Zugehörigkeit der Sequenzen in den Clusterbäumen nachvollziehbar ist. Mit Hilfe dieser Datenbank können z.B. Informationen über Sequenzanzahl, Clusternummer, EC-Nummer und E-Wert abgerufen werden.

In *tee* befinden sich Sequenzen, die Aufteilung der Sequenzen in Sequenzabschnitte und die Zugehörigkeit der Sequenzabschnitte in den verschiedenen Clustern bei E-Werten zwischen 10^{-2} und 10^{-181}. Diese Sequenzen wurden verwendet, um über lokale BLAST Alignments die Domänenstruktur der Enzyme zu identifizieren. Auf diese Weise sind 2108181 Subsequenzen entstanden, die in die Datenbank eingetragen wurden. Da jede Domäne bei jedem E-Wert einem Cluster zugeordnet wurde, befinden sich in *tee,* da 180 E-Werte vorhanden sind, fast 379,5 Millionen Einträge, die Auskunft über die Clusterzugehörigkeit einzelner Sequenzen geben.

Da die Sequenzen beginnend bei E-Wert 10^{-181} und aufsteigend bis E-Wert 10^{-2} nach dem *single linkage* Verfahren geclustert wurden, existiert zu Beginn der Clusterung, bei E-Wert 10^{-181}, die größte Anzahl Cluster. Diese Cluster enthalten nur sehr wenige Sequenzen; meist bestehen diese Cluster aus einer Sequenz. Bei jedem E-Wert vereinigen sich Cluster, so dass die Clusteranzahl mit fortschreitender Clusterung abnimmt und sich die Sequenzanzahl in den Clustern vergrößert. Das Diagramm 3-1 zeigt den Verlauf der Clusterung in Abhängigkeit der verwendeten E-Werte.

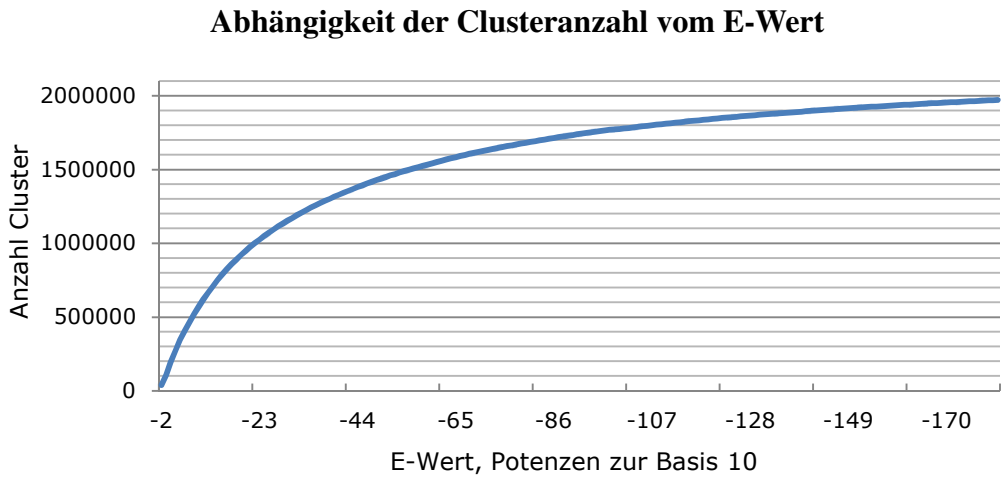


Diagramm 3-1: Clusteranzahl im Verlauf der Clusterung. Mit steigendem E-Wert nimmt die Clusteranzahl ab.

Bei einem E-Wert von 10^{-181} existieren 1971956 Cluster, bei E-Wert 10^{-2} sind 38691 Cluster vorhanden. Damit existieren 38691 Clusterbäume. Ausgehend vom Start der Clusterung bei

E-Wert 10^{-181}, werden die Sequenzen bis zum E-Wert 10^{-2} immer stärker geclustert. Wie in Diagramm 3-1 zu erkennen ist, nimmt die Clusteranzahl im Verlauf der Clusterung nicht linear ab, sondern die Clusterungsgeschwindigkeit nimmt zu, je größer der E-Wert wird.

3.1.3 Erhaltene EC-Kombinationen

Bei kleinen E-Werten werden Sequenzen geclustert, die sehr ähnlich zueinander sind. Bezogen auf das EC-System der Enzyme wird vermutet, dass sich bei kleinen E-Werten auch Sequenzen gruppieren, die der gleichen EC-Klasse angehören. Steigt der E-Wert an, ist zu erwarten, dass zunehmend Sequenzen geclustert werden, die aus verschiedenen EC-Klassen stammen. Die folgende Tabelle zeigt Einzelheiten zu der EC Zusammensetzung der Cluster bei verschiedenen E-Werten. Es wurden ausschließlich Cluster berücksichtigt, die unterschiedliche EC-Nummern enthalten. Diese Angabe wurde in der Spalte „Anzahl Cluster“ eingetragen. Bei jedem E-Wert wurden alle unterschiedlichen EC-Nummern eines Clusters miteinander verglichen; der jeweils größte Unterschied der EC-Nummern steht in der Tabelle. Sind zum Beispiel drei EC-Nummern in einem Cluster vorhanden und unterscheidet sich die erste EC-Nummer von der zweiten anhand ihrer Subklasse, wird dieser Wert nicht in die Tabelle eingetragen, falls ein EC-Klassenunterschied zwischen anderen EC-Nummern dieses Clusters vorhanden ist. In diesem Fall erhält ausschließlich die Spalte „Unterschiedliche Klasse“ einen Eintrag.

E-Wert	Anzahl Cluster	Unterschiedliche Klasse	Unterschiedliche Subklasse	Unterschiedliche Subsubklasse	Unterschiedliche Seriennummer
10^{-2}	14596	2991	2146	4729	4730
10^{-20}	1713	412	209	653	439
10^{-40}	1078	241	111	391	335
10^{-60}	842	183	75	317	267
10^{-80}	566	146	50	232	138
10^{-120}	369	67	35	130	137
10^{-140}	248	51	18	96	83
10^{-160}	182	43	18	66	55
10^{-181}	149	25	15	56	53

Tabelle 3-1: Erhaltene EC-Kombinationen der Clusterung für verschiedene E-Werte.

Vergleicht man den Anteil der in den Clustern enthaltenen EC-Klassen mit allen Vergleichen pro E-Wert, ist leicht zu erkennen, dass mit ansteigendem E-Wert der Anteil verschiedener EC-Klassen in den Clustern zunimmt. Bei einem E-Wert von 10^{-181} beträgt der Anteil an

unterschiedlichen Klassen in den Clustern 16,8%, bei E-Wert 10^{-2} sind in 20,5% von 14596 Clustern verschiedene EC-Nummern enthalten, die zu unterschiedlichen EC-Klassen angehören. Von E-Wert 10^{-20} (412 unterschiedliche EC-Klassen) bis 10^{-2} (2991 unterschiedliche EC-Klassen) nimmt die Zahl unterschiedlicher EC-Nummern in den Clustern sprunghaft zu. Der Grund dafür ist der sprunghafte gestiegene Anteil der Cluster, die bei jeweiligen E-Werten entstanden sind (1713 Cluster bei E-Wert 10^{-20}, 14596 bei E-Wert 10^{-2}).

3.1.4 Sequenzen mit mehr als einer EC-Nummer

In der Datenbank sind alle Sequenzen enthalten, die eine vollständige EC-Nummer besitzen. Auch Sequenzen, die pro Sequenz mehr als eine EC-Nummer besitzen, sind in der Datenbank enthalten. Das Diagramm 3-2 zeigt die Anzahl aller Enzymsequenzen der Datenbank *seq* und die Anzahl der jeweils zugewiesenen EC-Nummern pro Sequenz. Die größte Gruppe der Sequenzen mit mehr als eine EC-Nummer bildet die Gruppe der Sequenzen, die je Sequenz zwei EC-Nummern besitzen. Wie dem Diagramm zu entnehmen ist, besitzen die meisten Sequenzen eine EC-Nummer, es sind aber auch fünf Sequenzen vorhanden, denen acht EC-Nummern zugeordnet wurden. Diese Sequenzen haben eine Länge von etwa 3000 Aminosäuren und gehören zur Gruppe der Fettsäuren Synthasen. Als Beispiel sei hier das Enzym „Fatty acid synthase“, SWISS-PROT Accession-Nummer Q71SP7, genannt, das acht EC-Nummern aus vier EC-Klassen enthält.

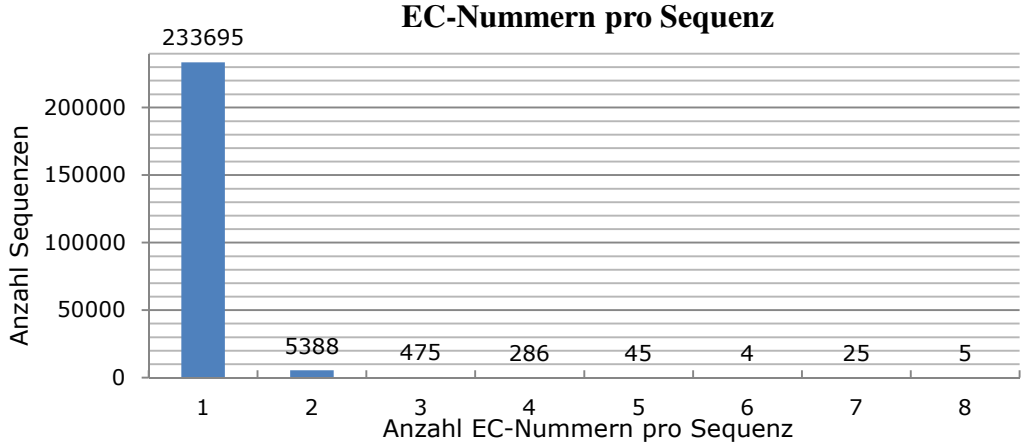


Diagramm 3-2: Übersicht über die Anzahl Sequenzen und der Anzahl ihrer EC-Nummern.

3.1.5 Verteilung der EC-Klassen

In der Datenbank sind Sequenzen aller sechs EC-Klassen vorhanden. Abhängig von der Anzahl der in den Datenbanken SWISS-PROT und TrEMBL enthaltenen Enzymsequenzen, variiert der Anteil der in der Datenbank enthaltenen EC-Klassen. Wie anhand des Diagramms 3-3 zu erkennen ist, enthält die Datenbank zu 29,4% Hydrolasen. Zwei weitere große EC-Klassen der Datenbank bilden die Klassen der Transferasen und der Oxidoreduktasen, die zusammen einen 52,7 prozentigen Anteil an der Datenbank haben. Die restlichen drei EC-Hauptklassen der Lyasen, Ligasen und Isomerasen bilden mit zusammen 17,8% einen verhältnismäßig geringen Anteil der Datenbank.

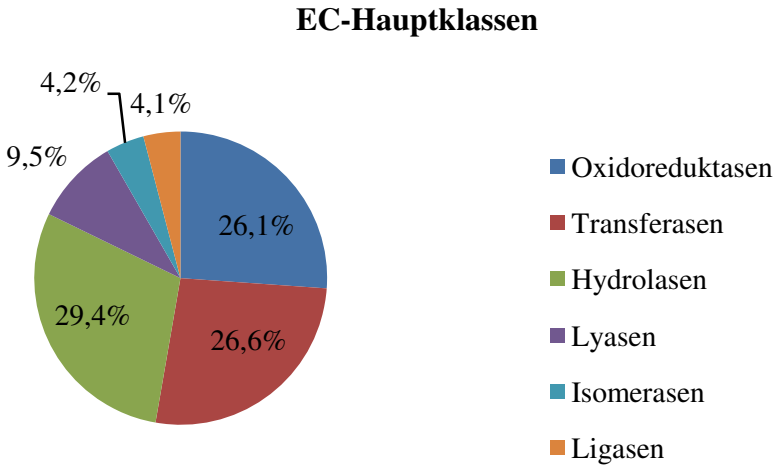


Diagramm 3-3: Verteilung der EC-Hauptkassen, der in der Datenbank *tee* enthaltenen Enzyme.

3.1.6 Cluster mit mindestens zehn Sequenzen (und > 1 EC-Nummer)

In der Datenbank existieren bei E-Wert 10^{-2} 38691 Cluster (vgl. Abschnitt 3.1.2). 3843 Cluster enthalten bei E-Wert 10^{-2} mindestens zehn Sequenzen. 2199 von diesen Clustern enthalten mehr als eine EC-Nummer. Eine Übersicht der Cluster, die mindestens zehn Sequenzen, bzw. zusätzlich mehr als eine EC-Nummer enthalten, ist im Diagramm 3-4 für verschiedene E-Werte zusammengefasst.

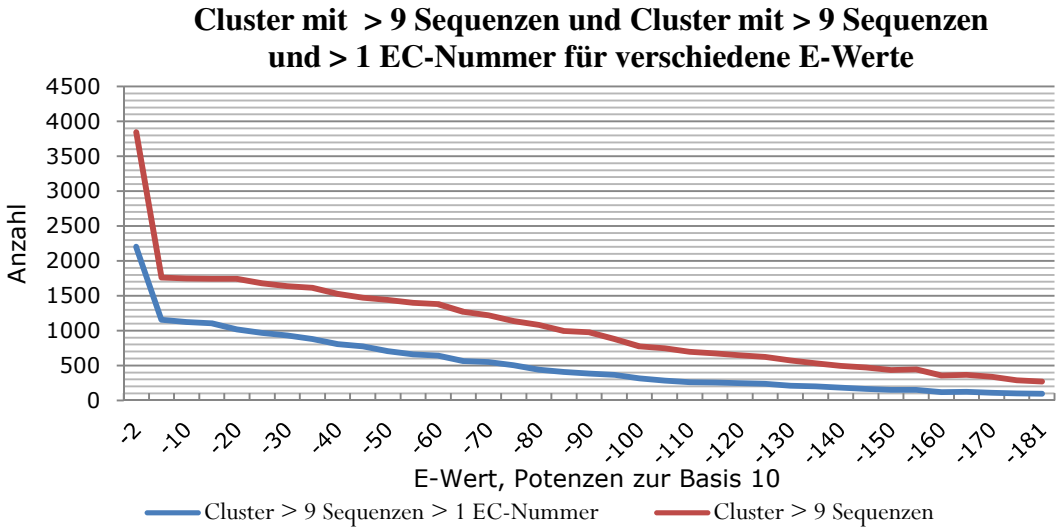


Diagramm 3-4: Anteil der Cluster mit > 9 Sequenzen und Cluster mit > 9 Sequenzen und > 1 EC-Nummer in Abhängigkeit vom E-Wert.

Der Anteil der Cluster, die mindestens zehn Sequenzen enthalten, ist bei E-Wert 10^{-2} besonders hoch. Die Anzahl dieser Cluster steigt von 1759 Cluster bei E-Wert 10^{-3} auf 3843 Cluster bei E-Wert 10^{-2}. Dies bedeutet eine Zunahme dieser Cluster um 45,8%. Die Anzahl der Cluster, die mindesten zehn Sequenzen und mehr als eine EC-Nummer enthalten, nimmt von E-Wert 10^{-3} bis E-Wert 10^{-2} um 52,4% zu. Vor diesem großen Anstieg verläuft die Clusterung bis hierhin relativ konstant.

3.1.7 Cluster mit mehr als zehn EC-Nummern in Abhängigkeit vom E-Wert

Im Verlauf der Clusterung werden Sequenzen mit unterschiedlichen EC-Nummern geclustert. Abhängig vom E-Wert, nimmt die Anzahl der Sequenzen in den Clustern zu (vgl. Abschnitt 3.1.2). Im Hinblick auf das EC-System, nimmt mit steigendem E-Wert auch die Anzahl unterschiedlicher EC-Nummern in den Clustern zu (vgl. Tabelle 3-1). In Abhängigkeit der Ähnlichkeit der Sequenzen, kann die Anzahl unterschiedlicher EC-Nummern in einem Cluster besonders groß sein. Das folgende Diagramm zeigt die Anzahl der Cluster, die im Verlauf des Clusterverfahrens mehr als zehn EC-Nummern enthalten.

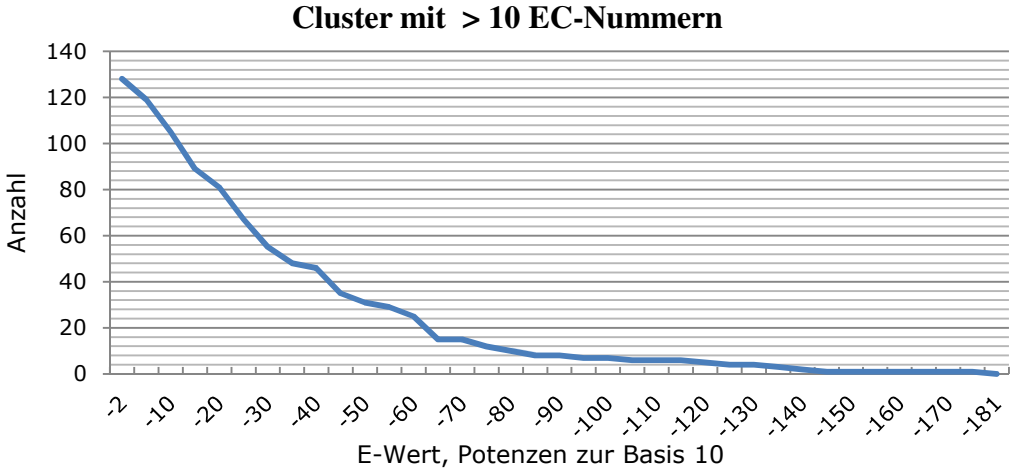


Diagramm 3-5: Verteilung von Clustern, die bei verschiedenen E-Werten mehr als zehn EC-Nummern enthalten.

Zu Beginn der Clusterung, bei E-Wert 10^{-181}, sind erwartungsgemäß keine Cluster vorhanden, die mehr als zehn EC-Nummern enthalten. Dies liegt zum einen an der geringen Anzahl Sequenzen, die bei diesem E-Wert geclustert wurden, zum anderen werden bei diesem E-Wert Sequenzen geclustert, die sehr ähnlich zueinander sind.

3.1.8 Anzahl EC-Nummern in Clustern in Abhängigkeit vom E-Wert

Da sich beginnend bei E-Wert 10^{-181} bis E-Wert 10^{-2} Cluster vereinigen, die Clusteranzahl dadurch ab- und die Sequenzanzahl innerhalb der Cluster zunimmt (vgl. Abschnitt 3.1.2), ist die Wahrscheinlichkeit, dass neue EC-Nummern zu einem Cluster hinzukommen, groß. In Diagramm 3-6 ist die durchschnittliche Anzahl EC-Nummern der Cluster gegen die untersuchten E-Werte aufgetragen. Dabei wurden für jeden vorhandenen E-Wert alle EC-Nummern jeden Clusters addiert und durch die vorhandene Clusteranzahl bei diesem E-Wert dividiert.

Beginnend bei E-Wert 10^{-181} nimmt die durchschnittliche Anzahl der in den Clustern enthaltenen EC-Nummern zunächst leicht zu. Bei einem E-Wert von etwa 10^{-50} nimmt die Anzahl der in den Clustern enthaltenen EC-Nummern leicht ab. Bei E-Wert 10^{-2} steigt der Anteil der in den Clustern enthaltenen EC-Nummern sprunghaft auf etwa 1,7 EC-Nummern je Cluster an.

Das Diagramm ist im Zusammenhang mit den Diagrammen 3-1 und 3-4 zu sehen. Wie das Diagramm 3-1 zeigt, nimmt die Anzahl der Cluster bei E-Wert 10^{-2} besonders stark ab. Gleichzeitig nimmt besonders bei E-Wert 10^{-2} die Anzahl der Sequenzen in den Clustern besonders stark zu (vgl. Diagramm 3-4). Besonders Sequenzen, die bei E-Wert 10^{-2} geclustert werden, besitzen unterschiedliche EC-Nummern. Dabei zeigen die Diagramme 3-4 und 3-6, dass ein Sprung bei E-Wert 10^{-2} existiert.

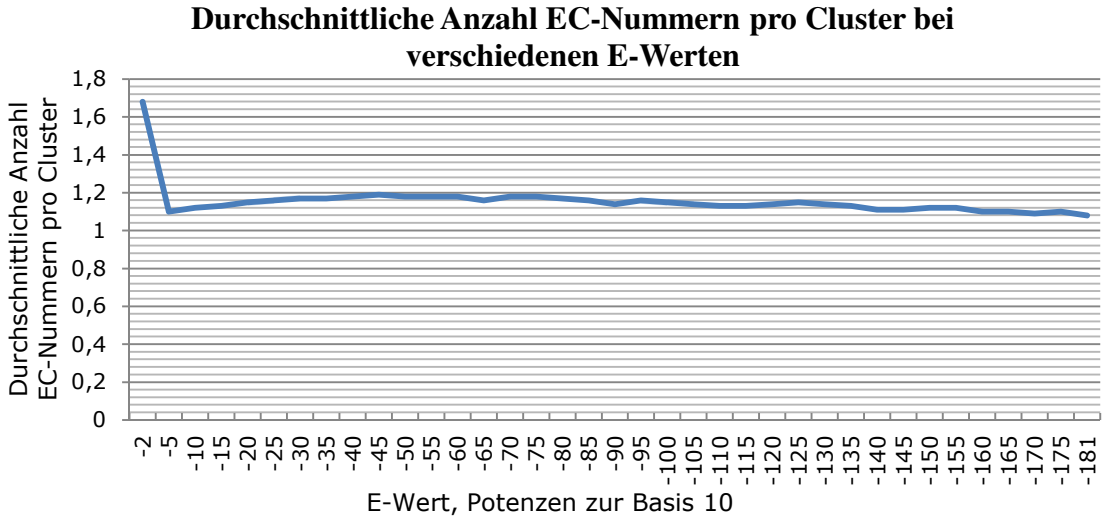


Diagramm 3-6: Durchschnittliche Anzahl EC-Nummern in den Clustern in Abhängigkeit vom E-Wert.

3.1.9 Sequenzanzahl in Clustern bei E-Wert 10^{-2}

Im Diagramm 3-7 wird die Sequenzanzahl der Cluster bei E-Wert 10^{-2} dargestellt. Die Angaben beziehen sich auf alle Clusterbäume der Datenbank. 1609 Cluster enthalten 10 bis 19 Sequenzen, 506 Cluster enthalten 20 bis 29 Sequenzen. Der weitaus größte Teil der Cluster ist bei E-Wert 10^{-2} sehr klein. 34536 Cluster bestehen aus 1 bis 9 Sequenzen. Dies entspricht bei E-Wert 10^{-2} einem Anteil von 89,3% aller Clusterbäume. 343 Cluster enthalten zwischen 550 bis 1000 Sequenzen. Diese extremen Werte wurden, um die Lesbarkeit des Diagramms zu erhalten, nicht in das Diagramm eingetragen. Insgesamt bestehen 3843 Cluster, bzw. 9,9% aller Cluster bei E-Wert 10^{-2} aus mindestens zehn Sequenzen. Es muss in diesem Zusammenhang erwähnt werden, dass 197 Cluster existieren, die bei E-Wert 10^{-2} zwischen 1000 und 10000 Sequenzen groß sind. 13 Cluster enthalten bei E-Wert 10^{-2} mehr als 10000 Sequenzen. Das größte Cluster, das bei E-Wert 10^{-2} existiert, enthält 991205 Sequenzen.

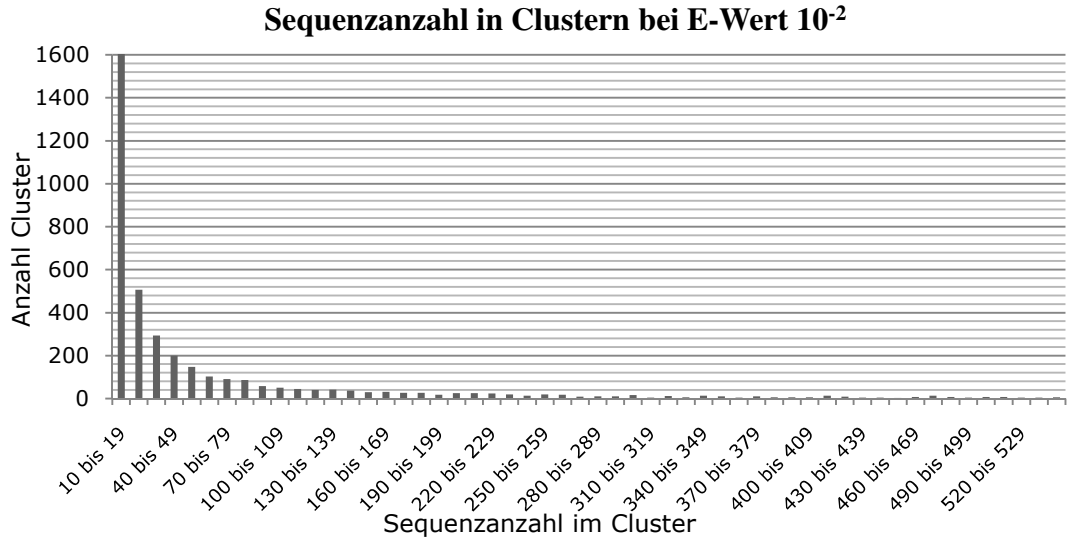


Diagramm 3-7: Sequenzanzahl in Clustern bei E-Wert 10^{-2}.

3.1.10 Überblick über einige Clusterbäume

Es ist sehr schwierig, einen zusammenhängenden Überblick über die Muster, EC-Nummern und Sequenzzahlen in mehreren zehntausend Clusterbäumen zu liefern. Die bisherigen Diagramme zeigen die Analyse vieler Werte im Einzelnen. Das Diagramm 3-8 zeigt einen Ausschnitt von 100 Clustern bei E-Wert 10^{-2}, bei denen die EC-Nummern und Sequenzanzahlen in einem Diagramm zusammengefasst wurden. Das Diagramm zeigt exemplarisch für den kompletten Datensatz einige willkürlich gewählte Cluster der Datenbank, um einen Eindruck über die Zusammensetzung der Clusterbäume der Datenbank zu liefern.

Viele der dargestellten Clusterbäume sind relativ klein. Sie enthalten jeweils etwa fünf Sequenzen. Drei Cluster besitzen mehr als 50 Sequenzen. Die Anzahl der EC-Nummern in den Clustern variiert stark. Obwohl drei Clusterbäume mehr als 50 Sequenzen enthalten, zeigt besonders die Auswahl der drei größten Cluster dieses Diagramms, dass die Anzahl der EC-Nummern in einem Cluster nicht ausschließlich von der Anzahl der Sequenzen im Cluster abhängig ist, wie anhand der Diagramme 3-4 und 3-6 zu vermuten wäre.

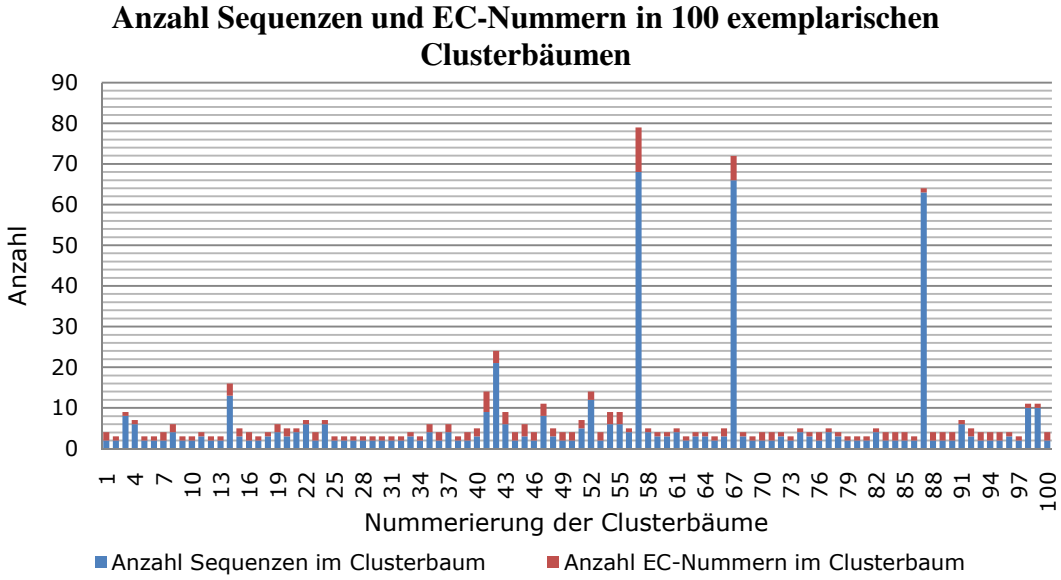


Diagramm 3-8: Überblick über 100 exemplarische Clusterbäume und ihrer Sequenz- und EC-Zusammensetzung.

3.2 Erstellung der Sequenzmuster

3.2.1 Rechenzeitbedarf

Die Erstellung der Sequenzmuster bedarf mehrerer Schritte, von denen einige besonders zeitintensiv sind.

Besonders rechenintensiv sind die Berechnungen der Programme CLUSTAL W bei der Erstellung globaler Alignments (vgl. Abschnitt 2.5.2.2), sowie die Berechnungen des Programms *patmatdb*, das Datenbanken mit erstellten Sequenzmustern nach Treffern durchsucht (vgl. Abschnitt 2.13.1).

Wie schon bei der Clusterung der Sequenzen im Abschnitt 3.1.1 beschrieben wurde, ist der Zeitbedarf der Mustererstellung stark abhängig von den zur Verfügung stehenden Computerressourcen und von den Größen der Datenbanken. Die Größen der Datenbanken TrEMBL und SWISS-PROT haben Einfluss auf die Menge der enthaltenen Enzymsequenzen und dadurch direkten Einfluss auf die Menge der zu erstellenden Muster. Auf einem Computercluster mit 32 CPUs (mit je 2000 MHz) lassen sich CLUSTAL W Alignments und die Suche nach Treffern mit *patmatdb* in ca. 4 Wochen durchführen.

3.2.2 Mustererstellung

Aus den Sequenzen aus 118947 Clustern der 38691 Clusterbäume (vgl. Abschnitt 3.1.2) wurden Alignments erstellt, weil sich die Anzahl der im Cluster enthaltenen EC-Nummern von Cluster zu Cluster in einem Clusterbaum änderte (vgl. Abschnitt 2.5.1). Die folgenden Abschnitte beschreiben die Eigenschaften der erstellten 118947 Sequenzmuster.

3.2.3 Bei welchen E-Werten wurden Muster generiert?

Das Diagramm 3-9 zeigt die Verteilung und Häufigkeit der erstellten Muster in Abhängigkeit vom E-Wert. Wie dem Diagramm zu entnehmen ist, wurden hauptsächlich bei hohen E-Werten Muster erstellt. Da an der Spitze der Clusterbäume bei E-Wert 10^{-2} aus den Sequenzen aller Cluster Muster generiert wurden, sind diese Daten nicht im Diagramm eingetragen worden. Mit ansteigendem E-Wert nimmt die Anzahl der generierten Muster mit leichten Schwankungen grundsätzlich zu. Bei einem E-Wert von 10^{-5} wurden 3173 Muster generiert. Abhängig von der Clusterungsgeschwindigkeit, siehe Diagramm 3-1, nimmt die Anzahl der erstellten Muster mit steigendem E-Wert zu.

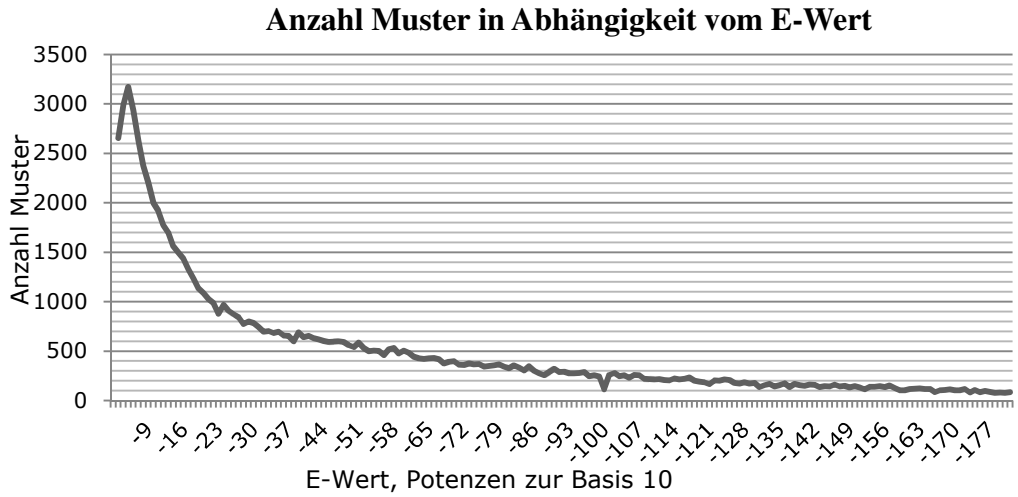


Diagramm 3-9: Die Anzahl der generierten Muster ist vom E-Wert abhängig, da bei ansteigendem E-Wert zunehmend die Kriterien zur Mustererstellung erfüllt werden.

3.2.4 Mustergenerierung und Sequenzanzahl

Im Diagramm 3-10 wird dargestellt, wie viele Sequenzen in den Clustern vorhanden waren, aus deren Sequenzen Muster erstellt wurden. Die meisten Muster wurden aus Cluster generiert, die zwischen zwei und neun Sequenzen bestehen. Mehr als 80000 Muster wurden aus weniger als zehn Sequenzen pro Muster generiert. Dies ist der Tatsache geschuldet, dass sehr viele Cluster aus sehr wenigen Sequenzen bestehen. Auch bei E-Wert 10^{-2} bestehen die meisten Cluster aus weniger als zehn Sequenzen (vgl. Diagramm 3-7). Bei 10 bis 19 Sequenzen pro Muster, hat die Musterzahl abgenommen. Im weiteren Verlauf nimmt die Anzahl der erstellten Muster kontinuierlich ab, je mehr Sequenzen im Cluster enthalten sind. Die Abhängigkeit der erstellten Muster von der Anzahl der zu Grunde liegenden Sequenzen eines Clusters korreliert direkt mit der Clusterungsgeschwindigkeit (vgl. Diagramm 3-1) und der Anzahl unterschiedlicher EC-Nummern in einem Cluster in Abhängigkeit des E-Wertes (siehe Diagramm 3-6). Ist die Anzahl der Sequenzen eines Clusters größer 1000, wurde aus diesen Sequenzen kein globales Alignment und damit kein Muster erstellt (vgl. Abschnitt 2.6).

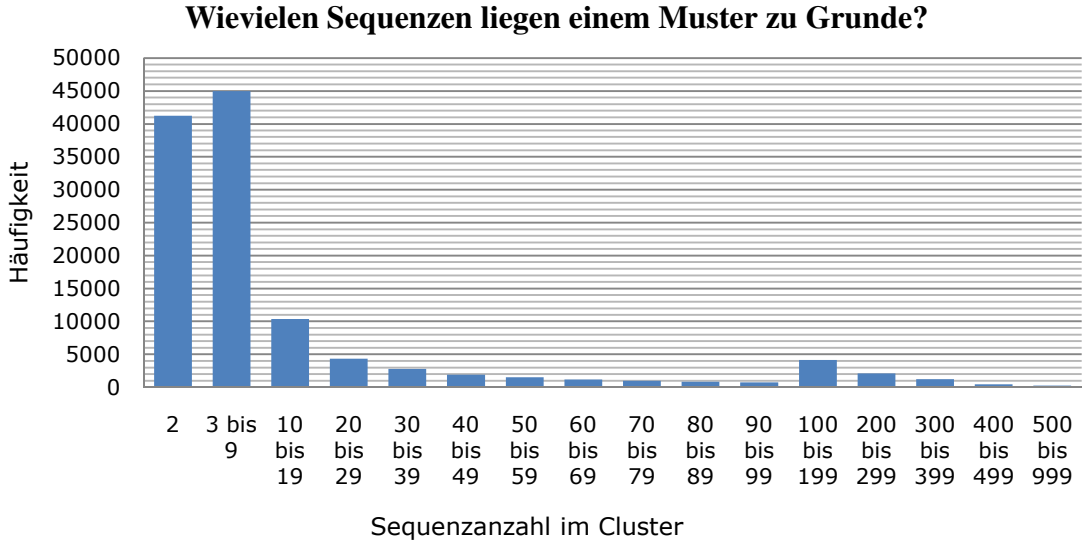


Diagramm 3-10: Generierung von Mustern in Abhängigkeit der Sequenzanzahl im Cluster.

3.2.5 Musterlänge: längste und kürzeste Muster

Im Diagramm 3-11 wird die Verteilung der Länge der in der Datenbank gespeicherten Muster dargestellt. Wie man leicht erkennen kann, sind die meisten Muster zwischen 9 und 200 Musterpositionen lang. Dieser Bereich wurde im Diagramm in zehn Teilbereiche aufgeteilt. Die Mindestlänge, der in der Datenbank eingetragenen Muster ist 8 Musterpositionen. 567

Muster haben exakt diese Länge. Kürzere Muster wurden aufgrund des damit verbundenen Verlusts der biologischen Signifikanz aus der Datenbank gelöscht (vgl. Abschnitt 2.7). Der Bereich der Musterlänge zwischen neun bis 100 Musterpositionen ist recht homogen. Jeweils etwa 5000 Muster verteilen sich in diesem Diagrammabschnitt auf die neun Bereiche zwischen neun und 100 Musterpositionen. Beginnend bei Mustern mit bis zu 200 Musterpositionen, nimmt die Zahl der Muster ab, je länger das Muster wird. 24768 Muster sind zwischen 201 und 1000 Musterpositionen lang. 513 Muster sind länger als 1000 Musterpositionen, wobei davon zwei Muster länger sind als 4000 Positionen.

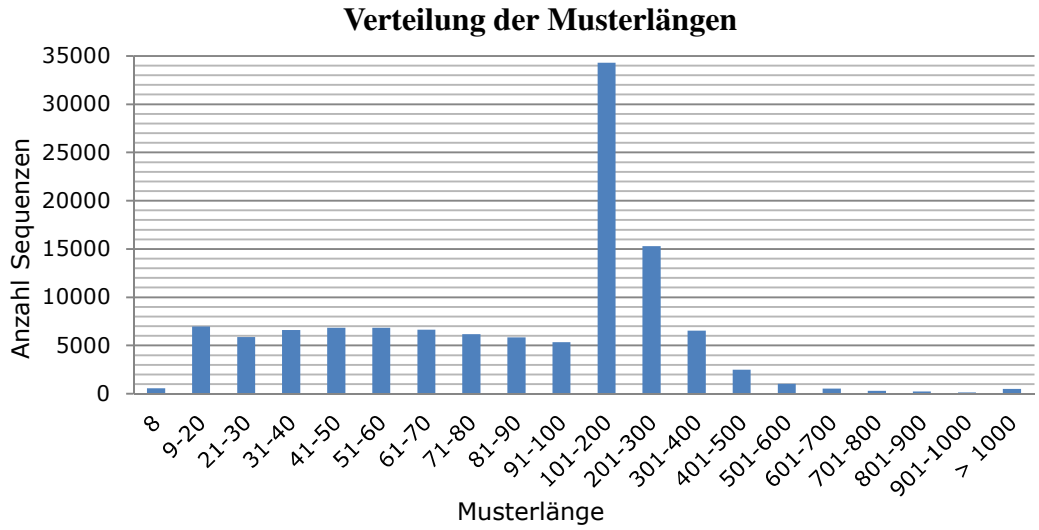


Diagramm 3-11: Verteilung der Musterlängen der in der Datenbank gespeicherten Sequenzmuster.

3.2.5.1 Muster mit mehr als 4000 Positionen

In der Datenbank sind zwei Muster vorhanden, die jeweils länger als 4000 Positionen sind. Das erste Muster wurde aus den Sequenzen Q0S5H6, Sequenzabschnitt 74 bis 5236 und Q0S5D8, Sequenzabschnitt 1 bis 5182, E-Wert 10^{-2}, generiert. Die Enzyme gehören zu der Gruppe der Nicht-Ribosomalen Peptid Synthetasen an. Das aus diesen Sequenzen generierte Muster ist 4062 Positionen lang. Im zweiten Cluster befinden sich die Sequenz Q8NEZ4, Sequenzabschnitt 1 bis 4911 und die Sequenz Q8BRH4, Sequenzabschnitt 1 bis 4903, E-Wert 10^{-2}. Die Enzyme tragen die EC-Nummer 2.1.1.43. Das aus diesen Sequenzen generierte Muster ist 4498 Positionen lang.

3.2.6 Musterlänge in Abhängigkeit der Sequenzanzahl

Es wurde vermutet, dass ein Alignment aus wenigen Sequenzen, zu einem längeren Muster führt, bzw. je mehr Sequenzen in einem Alignment vorhanden sind, desto kürzer würde das Sequenzmuster werden. Je mehr Sequenzen in einem Alignment vorhanden sind, desto schwieriger ist es für CLUSTAL W, alle konservierten Positionen zu finden. Außerdem steigt die Gefahr bei sehr vielen Sequenzen, dass Sequenzen im Alignment vorkommen, die wesentlich länger oder kürzer sind, als andere im Alignment vorhandene Sequenzen. Das führt dazu, dass die Anzahl der konservierten Positionen im Muster mit steigender Sequenzanzahl im Alignment abnimmt. Das Diagramm 3-12 zeigt die Sequenzmusterlänge in Abhängigkeit der zu Grunde liegenden Sequenzen. Die Daten für das Diagramm wurden aus der Datenbank für alle Muster bei E-Wert 10^{-2} gewonnen.

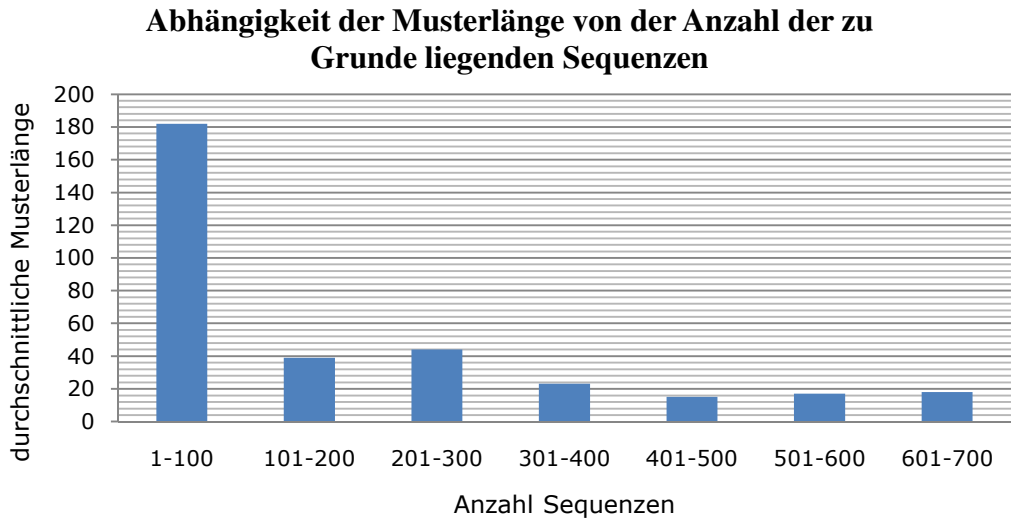


Diagramm 3-12: Die Musterlänge ist abhängig von den zu Grunde liegenden Sequenzen.

Dem Diagramm 3-10 ist deutlich zu entnehmen, dass die erstellten Muster mit steigender Sequenzzahl im Alignment kürzer werden. Ein Muster, das auf Grundlage von bis zu 100 Sequenzen erstellt wurde, ist beispielsweise durchschnittlich 182 Positionen lang. Ein Muster, das auf Grundlage von 601-700 Sequenzen erstellt wurde, besteht aus durchschnittlich 18 Positionen.

3.2.7 Muster für Sequenzen mit einer EC-Nummer

In der Datenbank sind Muster für Sequenzen vorhanden, die eine EC-Nummer pro Sequenz tragen. Es existieren aber auch Sequenzen, die pro Sequenz mehrere EC-Nummern besitzen (vgl. Diagramm 3-2).

Das Diagramm 3-13 zeigt die Anzahl der Sequenzmuster mit einer EC-Nummer und die Verteilung dieser EC-Nummern auf die EC-Klassen.

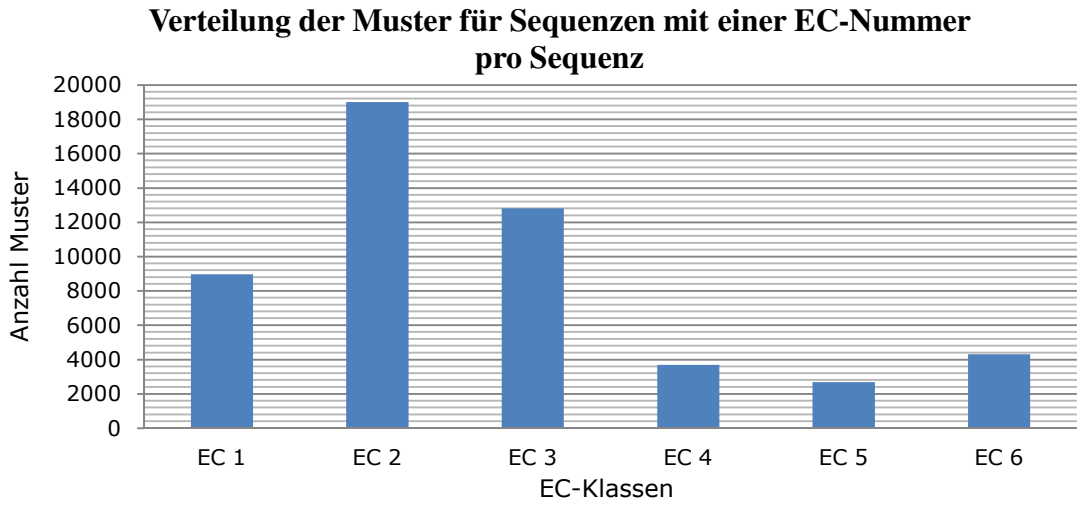


Diagramm 3-13: Erstellte Muster mit jeweils einer EC-Nummer und die Verteilung dieser Muster auf die Klassen des EC-Systems.

Insgesamt gibt es 51486 Muster, die jeweils nur eine EC-Nummer repräsentieren. Die größte Gruppe der Muster für Sequenzen mit einer EC-Nummer pro Sequenz bildet die Gruppe der Transferasen, EC-Klasse 2, die wenigsten Muster sind für die Gruppe der Isomerasen, EC-Klasse 5, vorhanden.

3.2.8 Untersuchung der Richtig-Positiven und Falsch-Positiven Treffer

Zunächst muss noch einmal festgehalten werden, dass aus Prinzip der Mustererstellung natürlich mindestens die Sequenzen mit dem speziellen Suchmuster gefunden werden müssen, aus denen das Muster erstellt wurde. Das ist für alle Muster der Datenbank der Fall. Zur Erklärung, wie ein Richtig-Positiver, oder Falsch-Positiver Treffer definiert wird, wird auf den Abschnitt 2.9 im Teil Daten, Algorithmen und Methoden verwiesen.

3.2.9 Einschätzung der Qualität der Muster anhand Richtig-Positiver und Falsch-Positiver Treffer

Von 118947 Sequenzmustern, die in der Datenbank enthalten sind, trafen 94329 Sequenzmuster ausschließlich die Sequenzen, auf deren Grundlage das Muster erstellt wurde. Das heißt, dass die Muster so spezifisch sind, dass keine anderen Sequenzen getroffen wurden. Mit 79,3% entspricht dies dem größten Teil der Datenbank. 20,7% aller Sequenzmuster hatten mehr Richtig-Positive, als aufgrund der Sequenzanzahl, aus denen ein Muster erstellt wurde, erwartet wurde.

Bei 13883 Sequenzmustern ist die Anzahl der Sequenzmuster, aus denen das Muster erstellt wurde, identisch mit allen Richtig-Positiven Treffern und identisch mit allen nach Definition Falsch-Positiven Treffern. Demnach wurden exakt alle Sequenzen getroffen, aus denen ein Muster erstellt wurde, aber auch die gleiche Zahl Falsch-Positive Treffer wurden erzielt.

Die folgende Tabelle stellt die Anzahl und das Verhältnis von Richtig-Positiven und Falsch-Positiven Treffern aus allen Mustern der Datenbank dar.

Vergleich Treffer	**Anzahl**
Anzahl Richtig-Positive > Anzahl Falsch-Positive	97860
Anzahl Richtig-Positive = Anzahl Falsch-Positive	2568
Anzahl Richtig-Positive < Anzahl Falsch-Positive	18519

Tabelle 3-2: Anzahl Richtig-Positive, Falsch-Positive Treffer der Sequenzmuster der Datenbank.

Wie man der Tabelle 3-2 entnehmen kann, haben 97860 oder 82,3% aller Muster mehr Richtig-Positive Treffer als Falsch-Positive Treffer. Bei 21087 Mustern oder 17,7% ist die Anzahl der Richtig-Positiven Treffer identisch oder geringer als alle Falsch-Positiven Treffer dieser Muster.

3.2.10 Sequenzmuster ohne Falsch-Positive Treffer

Aufgrund der im vorherigen Abschnitt dargestellten Daten konnte gezeigt wurde, dass die Mehrheit aller Muster ausschließlich die Sequenzen treffen, aus denen sie erstellt wurden, d.h. diese Muster sind sehr spezifisch. Sie sind so spezifisch, dass neben den Sequenzen, aus denen

das Muster erstellt wurde, keine anderen Sequenzen getroffen wurden. 94329 Muster treffen ausschließlich die Sequenzen, aus denen das Muster erstellt wurde. Von diesen Mustern haben 57183 oder 60,6% keine Falsch-Positiven Treffer. 3075 Muster haben mehr Richtig-Positive Treffer, als die Anzahl der Sequenzen, aus denen das Muster erstellt wurde (Stammsequenzen) und haben keinen Falsch-Positiven Treffer. Dies entspricht 2,6% aller Muster der Datenbank. Die folgende Tabelle fasst die Ergebnisse zusammen.

Vergleich Treffer	**Anzahl**
Richtig-Positive = Stammsequenzen, Falsch-Positive = 0	57183
Richtig-Positive > Stammsequenzen, Falsch-Positive = 0	3075

Tabelle 3-3: Muster ohne Falsch-Positive Treffer.

3.2.11 Extremwerte

Eine Übersicht, in welchem Maße Muster sehr viele Richtig-Positive Treffer, dabei aber wenige oder keine Falsch-Positiven Treffer haben, ist schwer darstellbar, da die Anzahl der Treffer von vielen Faktoren beeinflusst wird. Wie bereits erwähnt wurde, ist die Trefferanzahl abhängig von der Anzahl der Sequenzen, die dem Muster zu Grunde liegen. Es soll aber in diesem Zusammenhang auf einige Extremwerte eingegangen werden.

Es existieren zwei Muster in der Datenbank, die aus mehr als 100 Sequenzen generiert wurden und mehr als 100 Richtig-Positive, aber keine Falsch-Positiven Treffer haben.

Das erste Muster besteht aus 37 Positionen. Es ist auf Grundlage von 100 Sequenzen entstanden, die bei einem E-Wert von 10^{-33} geclustert wurden. Alle Sequenzen stammen aus der Ursprungsdatenbank SWISS-PROT. Das Muster hat 101 Richtig-Positive Treffer; es wurde also eine Sequenz mehr getroffen, als schon anhand der Ursprungssequenzen zu erwarten war. Dieses Muster trägt die EC-Nummern EC 2.7.7.7, EC 2.7.7.49, EC 3.1.26.4, EC 3.4.23.16, EC 3.4.23.47, EC 3.4.23.- und EC 3.6.1.23. Die EC-Nummern gehören zu den EC-Klassen der Transferasen und Hydrolasen.

Das zweite Muster, das in diesem Zusammenhang erwähnt wird, wurde aus 131 Sequenzen generiert, hat 131 Richtig-Positive und keine Falsch-Positive Treffer. Das Muster ist 65 Positionen lang. Die Clustersequenzen haben einen E-Wert von 10^{-58}. Es ist eines von 12 Mustern in der Datenbank für die EC-Nummer 2.7.11.24. Die Sequenzen stammen alle, wie im vorherigen Beispiel, aus der Ursprungsdatenbank SWISS-PROT. Alle anderen Sequenzmuster

für die EC-Nummer 2.7.11.24 haben ebenfalls ausschließlich Richtig-Positive und keine Falsch-Positiven Treffer, wie die folgende Tabelle, ein Ausschnitt aus der Tabelle *veri*, zeigt.

E-Wert	Clusterbaum	Knoten	EC-Nummer	Sequenz-anzahl	Muster-länge	Richtig-Positive	Falsch-Positive
10^{-2}	5268	1	2.7.11.24	5	653	5	0
10^{-2}	5269	1	2.7.11.24	2	745	2	0
10^{-2}	5360	1	2.7.11.24	2	385	2	0
10^{-2}	5871	1	2.7.11.24	2	507	2	0
10^{-16}	6739	3	2.7.11.24	3	432	3	0
10^{-17}	300713	21408	2.7.11.24	7	327	7	0
10^{-31}	300713	33173	2.7.11.24	7	353	7	0
10^{-36}	300713	35083	2.7.11.24	4	376	4	0
10^{-40}	300713	36394	2.7.11.24	2	470	2	0
10^{-50}	300713	38825	2.7.11.24	3	222	3	0
10^{-55}	300713	40279	2.7.11.24	3	204	3	0
10^{-58}	300713	40977	2.7.11.24	131	65	131	0

Tabelle 3-4: Ausschnitt aus der Tabelle *veri*: Detailierte Informationen zu Sequenzmustern der EC-Nummer 2.7.11.24.

3.2.12 Untersuchung von Richtig-Positiven und Falsch-Positiven Treffern bei E-Wert 10^{-2}

Im folgenden Abschnitt wird untersucht, ob ein Zusammenhang zwischen dem Ausmaß von Richtig-Positiven und Falsch-Positiven Treffern und dem E-Wert besteht, bei dem Sequenzen geclustert und aus diesen im weiteren Verlauf Muster erstellt werden. Im dargestellten Diagramm 3-14 wird die Anzahl der Muster bei zehn E-Werten gezeigt und die durchschnittliche Anzahl Falsch-Positiver Treffer. Für die Berechnung des Falsch-Positiven Durchschnitts wird bei jedem E-Wert die Anzahl aller Falsch-Positiver Treffer durch die Anzahl aller Muster gerechnet. Das Ergebnis ist ein durchschnittlicher Wert für alle Falsch-Positiven Treffer pro Muster.

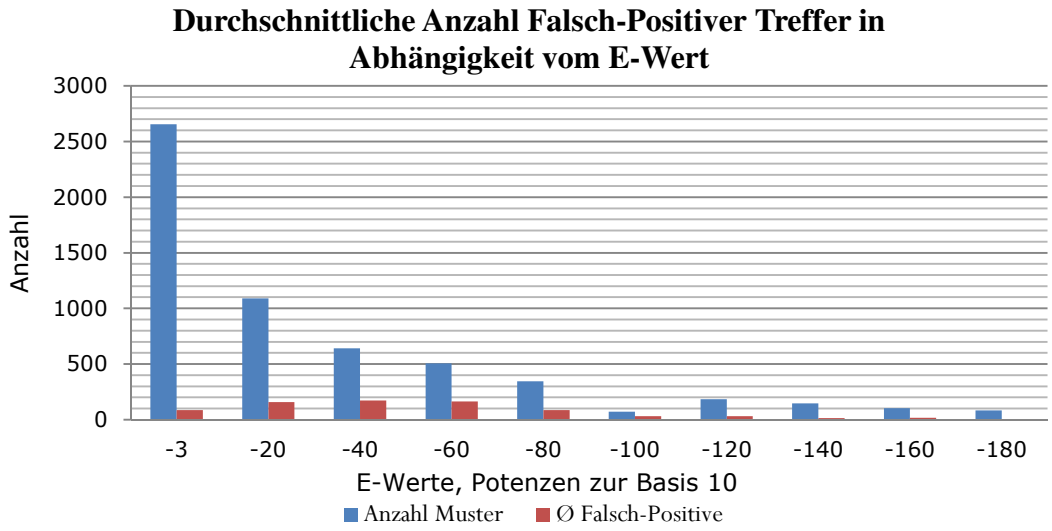


Diagramm 3-14: Durchschnittliche Anzahl Falsch-Positiver Treffern in Abhängigkeit vom E-Wert.

Von E-Wert 10^{-3} bis 10^{-40} steigt die Anzahl der durchschnittlichen Falsch-Positiven Treffer an, danach fällt die Anzahl kontinuierlich ab.

Das folgende Diagramm zeigt die durchschnittliche Anzahl von Richtig-Positiven Treffern in Abhängigkeit vom E-Wert. Die Daten wurden berechnet, indem bei jedem E-Wert alle Richtig-Positiven Treffer durch die Anzahl aller Muster geteilt wurden. Die Anzahl der den Mustern zu Grunde liegenden Muster wurde vor der Berechnung abgezogen, da die Sequenzen, die dem Muster zu Grunde liegen, immer getroffen werden.

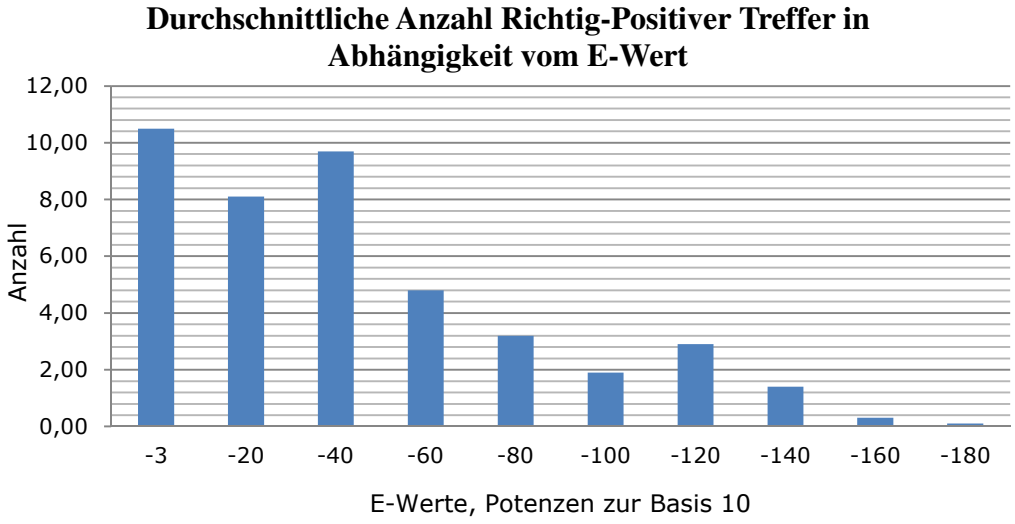


Diagramm 3-15: Durchschnittliche Anzahl Richtig-Positiver Treffern in Abhängigkeit vom E-Wert.

3.2.13 Abhängigkeit der Verteilung von Richtig-Positiven und Falsch-Positiven Treffern von der Musterlänge

Im folgenden Abschnitt wird untersucht, ob ein Zusammenhang zwischen dem Ausmaß von Richtig-Positiven und Falsch-Positiven Treffern und der Länge der Sequenzmuster existiert. Ein Zusammenhang ist wahrscheinlich, denn je kürzer ein Sequenzmuster ist, desto mehr potentielle Richtig-Positive oder Falsch-Positive Treffer können mit diesem Muster erzielt werden (vgl. Abschnitt 2.8.2). Es muss aber auch beachtet werden, dass auch sehr kurze Muster biologisch signifikante Treffer erzielen können, da die Wahrscheinlichkeit eines Treffers nicht nur von der Sequenzlänge, sondern auch von Faktoren, wie variable Lücken oder statistische Häufigkeit der Aminosäuren im Muster abhängt. Die Diagramme 3-16 und 3-17 zeigen die durchschnittliche Anzahl Richtig-Positiver und die durchschnittliche Anzahl Falsch-Positiver Treffer in Abhängigkeit von der jeweiligen Musterlänge. Zur Berechnung der eingetragenen Werte wurden bei E-Wert 10^{-2} alle Sequenzmuster verwendet, die zwischen 8 und 100 Musterpositionen lang sind. Alle Richtig-Positiven Treffer wurden durch die Anzahl der Muster geteilt.

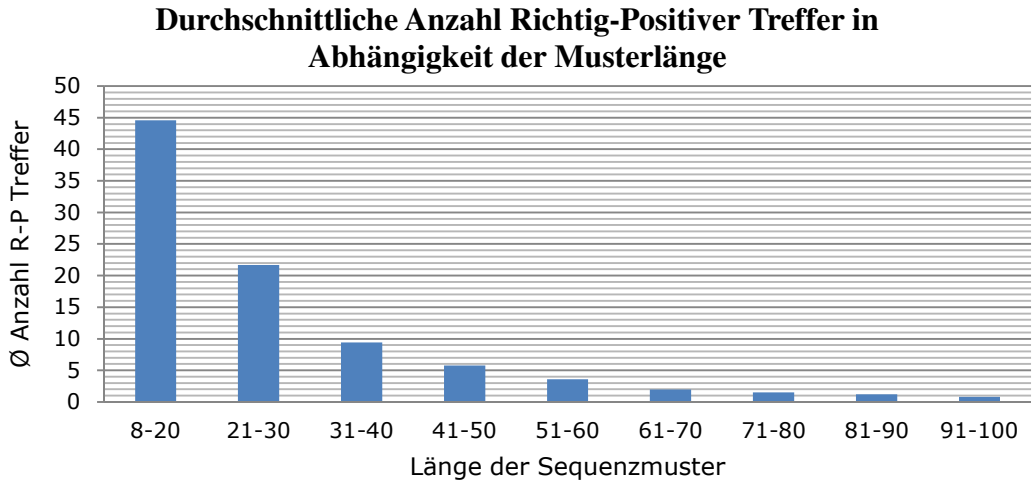


Diagramm 3-16: Durchschnittliche Anzahl Richtig-Positiver Treffer in Abhängigkeit der Musterlänge.

Ein Muster mit einer Länge von 8 bis 20 Musterpositionen hat durchschnittlich 45 Richtig-Positive Treffer erzielt, die über die Treffer der Sequenzen, die dem Muster zu Grunde liegen, hinausgehen. Je länger das Muster wird, desto weniger Richtig-Positive Treffer werden erzielt.

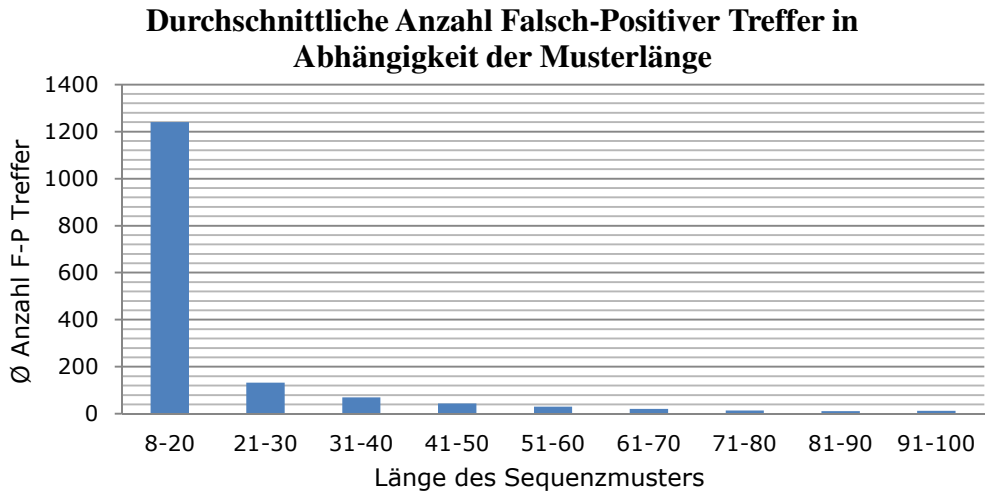


Diagramm 3-17: Durchschnittliche Anzahl Falsch-Positiver Treffer in Abhängigkeit der Musterlänge.

Wie schon die Untersuchung der Richtig-Positiven Treffer in Abhängigkeit der Musterlänge zeigt, ist auch bei der Untersuchung der durchschnittlichen Anzahl Falsch-Positiver Treffer zu sehen, dass je kürzer das Muster ist, desto mehr Treffer werden erzielt. Im vorliegenden Fall werden die getroffenen Sequenzen als Falsch-Positiv definiert. Ein Muster mit 8 bis 20 Musterpositionen hat durchschnittlich mehr als 1200 Falsch-Positive Treffer. Dieser extrem hohe Wert wird besonders durch 25 Muster hervorgerufen, die jeweils 8 Positionen lang sind und jeweils mehr als 19000 Falsch-Positive Treffer erzielten.

3.2.14 Abhängigkeit der Verteilung von Richtig-Positiven und Falsch-Positiven Treffern von der Ursprungsdatenbank

Die Sequenzen für die Clusteranalyse und der darauf folgenden Mustererstellung wurden den Sequenzdatenbanken SWISS-PROT und TrEMBL entnommen. Wurden nun Cluster erstellt, gibt es 3 Möglichkeiten, wie sich die Sequenzen der Datenbanken im Cluster verteilen:

- Die Sequenzen des Clusters stammen ausschließlich aus SWISS-PROT
- Die Sequenzen des Clusters stammen ausschließlich aus TrEMBL
- Die Sequenzen des Clusters stammen aus SWISS-PROT und TrEMBL

Stammen die Sequenzen eines Clusters, aus denen im weiteren Verlauf ein Muster erstellt wurde aus der Datenbank SWISS-PROT, wird nach Treffern des Musters in der Datenbank SWISS-PROT gesucht.

Stammen die Sequenzen eines Clusters, aus denen im weiteren Verlauf ein Muster erstellt wurde aus TrEMBL, wird nach Treffern des Musters in der Datenbank TrEMBL gesucht.

Stammen die Sequenzen eines Clusters, aus denen im weiteren Verlauf ein Muster erstellt wurde aus SWISS-PROT und TrEMBL, wird in beiden Datenbanken nach Treffer des Musters gesucht.

Diese Strategie ist nötig, da die Annotationen aus SWISS-PROT sehr viel verlässlicher sind, als die Annotationen aus TrEMBL, denn in der computerannotierten Datenbank TrEMBL können zusätzlich auch hypothetische Proteine oder nicht sicher zugeordnete Sequenzen enthalten sein. Da die Datenbank TrEMBL sehr viel mehr Proteine enthält, als die Datenbank SWISS-PROT, ist der Suchraum des Musters (regulären Ausdrucks) größer. Damit steigt die Wahrscheinlichkeit für Falsch-Positive Treffer, wenn ein Muster gegen TrEMBL gesucht wird, oder wenn nach Treffern eines Musters in SWISS-PROT und TrEMBL gesucht wird.

Das Diagramm 3-18 beschäftigt sich mit der Fragestellung, in welchem Ausmaß die Zugehörigkeit eines Sequenzmusters zu einer Datenbank, das Ausmaß von Falsch-Positiven Treffern beeinflusst. Die Anzahl Falsch-Positiver Treffer für 1000 Muster bis 100 Positionen bei E-Wert 10^{-2} bei der Suche von Mustern gegen die Datenbanken SWISS-PROT, TrEMBL und SWISS-PROT + TrEMBL wird im Diagramm dargestellt.

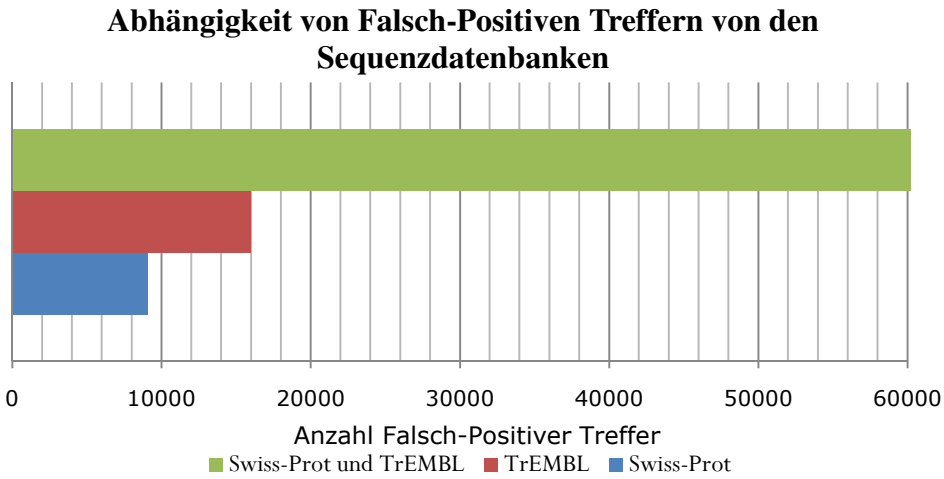


Diagramm 3-18: Abhängigkeit der Anzahl Falsch-Positiver Treffer von den Datenbanken, aus denen die Sequenzen für die Muster stammen.

Stammen die Sequenzen eines Musters aus der Datenbank TrEMBL, ist die Wahrscheinlichkeit, dass das resultierende Muster sehr viele Falsch-Positive Treffer erzielt hoch, da mit diesem Muster in der Datenbank TrEMBL gesucht wird. Die geringste Wahrscheinlichkeit, Falsch-Positive Treffer zu erzielen wird erreicht, wenn in der Datenbank

SWISS-PROT gesucht wird. Die wesentlich größte Wahrscheinlichkeit, Falsch-Positive Treffer zu erzielen, wird bei der Suche gegen SWISS-PROT und TrEMBL erreicht. Ein Muster, das auf Treffer in der Datenbank TrEMBL untersucht wird, erzielte bei den ausgewählten Mustern laut Diagramm 3-18 etwa 60 Falsch-Positive Treffer. Dagegen hat ein vergleichbares Muster, das auf Treffer gegen die Datenbank SWISS-PROT untersucht wird, etwa 8,5 Falsch-Positive Treffer.

3.3 Ergebnisse der Untersuchung ähnlicher Reaktionsmechanismen

Die Ergebnisse der Untersuchung der Edukt- und Produktmoleküle sowie der Co-Substrate der verglichenen chemischen Reaktionen von Enzymen auf eine größte gemeinsame Teilstruktur werden in diesem Abschnitt dargestellt. Bei der Untersuchung werden Reaktionen von Enzymen analysiert, die bei verschiedenen E-Werten geclustert wurden.

Mit zunehmendem E-Wert nimmt die Anzahl der EC-Kombinationen in den Cluster zu, da Cluster mit steigendem E-Wert zunehmend mehr Sequenzen enthalten. Von allen EC-Kombinationen werden die doppelt vorkommenden Kombinationen gelöscht. Für die Untersuchung der bei einer enzymatischen Reaktion beteiligten Moleküle mit dem c-MCS Algorithmus reduzieren sich die zu untersuchenden Reaktionen durch verschiedene Faktoren. Reaktionen, die nicht in der BRENDA-Datenbank vorhanden sind oder für die es keine Molfiles gibt, können nicht untersucht werden. Bei E-Wert 10^{-181} wurden 42 EC-Paarungen untersucht, bei 10^{-2} konnten 2462 Vergleiche untersucht werden. Einzelheiten zu weiteren Werten können der Tabelle 3-5 entnommen werden. Auf welche Weise die Auswahl der EC-Kombinationen getroffen wurde, kann im Abschnitt 2.10.5 nachgelesen werden.

E-Wert	**EC-Kombinationen gesamt**	**EC-Kombinationen ohne doppelte Kombinationen, ohne Kombinationen mit gleicher Seriennummer**	**Untersuchte Reaktionen für c-MCS**
10^{-2}	18628	3174	2462
10^{-40}	5098	603	476
10^{-80}	1212	258	230
10^{-120}	454	131	107
10^{-160}	228	71	55
10^{-181}	122	45	42

Tabelle 3-5: Übersicht der erhaltenen EC-Kombinationen bei verschiedenen E-Werten.

3.3.1 Untersuchung der Zusammensetzung der EC-Kombinationen für verschiedene E-Werte

Bei der Untersuchung der EC-Kombinationen wurden alle Vergleiche gelöscht, die sich lediglich anhand der vierten Stelle der EC-Nummer unterscheiden. Sind in einem Cluster EC-Nummern unterschiedlicher EC-Hierarchien vorhanden, wird in die Tabelle der jeweils größte Unterschied eingetragen. Sind in einem Cluster zum Beispiel die EC-Nummern EC 1.1.1.1, EC 2.2.2.2 und EC 1.2.3.4 vorhanden, liegt der größte Unterschied in der EC-Klasse zwischen EC 1.1.1.1 und EC 2.2.2.2. Nur dieser Wert wird gezählt und in der Tabelle berücksichtigt. Auf dieser Grundlage beruhen auch die Werte der Tabelle 3-7. Eine detailierte Übersicht über die Zusammensetzung der EC-Kombinationen bei verschiedenen E-Werten liefert die folgende Tabelle 3-6.

E-Wert	EC-Kombinationen gesamt	Unterschiedliche Klassen	Unterschiedliche Subklassen	Unterschiedliche Subsubklassen
10^{-2}	3174	1710	1240	224
10^{-40}	603	285	221	97
10^{-80}	258	120	86	52
10^{-120}	131	51	51	29
10^{-160}	71	32	19	20
10^{-181}	45	21	9	15

Tabelle 3-6: Übersicht über die EC-Zusammensetzung der erhaltenen EC-Kombinationen.

3.3.2 Untersuchung der Zusammensetzung der EC-Kombinationen für verschiedene E-Werte und Identität der bei den enzymatischen Reaktionen enthaltenen Moleküle

Die Zusammensetzungen der durch den c-MCS Algorithmus untersuchten Reaktionen der EC-Kombinationen sind der folgenden Tabelle 3-7 zu entnehmen. Bei E-Wert 10^{-181} stammen 17 EC-Nummern aus unterschiedlichen EC-Klassen und wurden miteinander verglichen. Bei E-Wert 10^{-2} steigt die Anzahl der paarweise verglichenen unterschiedlichen Klassen auf 774 an. Die größte Anzahl der verglichenen EC-Kombinationen wird bei E-Wert 10^{-2} erzielt. Bei diesem E-Wert wurden 2462 EC-Kombinationen miteinander verglichen. Dabei stammen 774 EC-Kombinationen aus unterschiedlichen Klassen, 706 EC-Kombinationen aus unterschiedlichen Subklassen und 982 Kombinationen aus unterschiedlichen Subsubklassen.

E-Wert	EC-Kombinationen für c-MCS	Unterschiedliche Klassen	Unterschiedliche Subklassen	Unterschiedliche Subsubklassen
10^{-2}	2462	774	706	982
10^{-40}	476	152	135	189
10^{-80}	230	80	58	92
10^{-120}	107	32	39	36
10^{-160}	55	20	11	24
10^{-181}	42	17	4	21

Tabelle 3-7: Übersicht über die EC-Zusammensetzungen der EC-Kombinationen, die mittels des c-MCS Algorithmus‘ untersucht wurden.

3.3.3 Prozentuale Auftragung der Untersuchung identischer Moleküle für EC-Kombinationen bei verschiedenen E-Werten

Bei verschiedenen E-Werten wurden EC-Kombinationen ermittelt, die auf dem Ergebnis der Clusteranalyse basieren. Sind in einem Cluster mehrere EC-Nummern vorhanden, werden die katalysierten Reaktionen dieser EC-Nummern untersucht (vgl. Abschnitt 2.10.5 und Tabelle 3-7), in welchem Maße ähnliche oder identische Moleküle bei den verglichenen Reaktionen vorhanden sind. Zu diesem Zweck wird der c-MCS Algorithmus genutzt, der die größte gemeinsame Teilstruktur der Moleküle ermittelt (vgl. Abschnitt 2.10). Das Diagramm 3-19 stellt bei verschiedenen E-Werten den prozentualen Anteil aller untersuchten Reaktionen dar, bei denen bei den verglichenen katalytischen Reaktionen mindestens ein Molekül in beiden Reaktionen identisch ist.

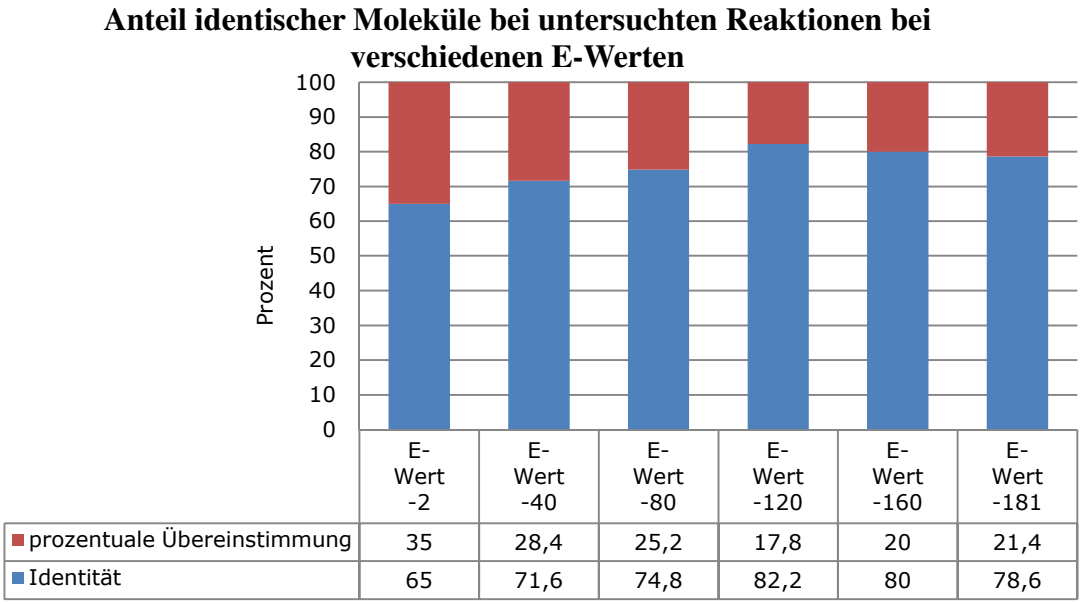


Diagramm 3-19: Ergebnis der Untersuchung identischer Moleküle bei verglichenen EC-Reaktionen. In der Zeile E-Wert werden die Potenzen zur Basis zehn angegeben.

Je kleiner der E-Wert ist, desto größer ist der Anteil der identischen Moleküle der bei den untersuchten Reaktionen beteiligten Moleküle. Bei E-Wert 10^{-181} sind in 78,6% aller Reaktionen die Edukt(e)/Produkt(e) oder Co-Substrat(e) identisch. Bis zum E-Wert 10^{-2} fällt dieser Anteil auf 65% ab.

3.3.4 Prozentuale Darstellung der prozentual größten gemeinsamen Teilstruktur der bei den paarweise verglichenen Reaktionen beteiligten Moleküle

Bei der Untersuchung der größten gemeinsamen Teilstruktur der bei den verglichenen Reaktionen beteiligten Moleküle mittels des c-MCS Algorithmus` werden nicht nur identische Moleküle identifiziert, sondern auch eine teilweise Übereinstimmung ist möglich. Die prozentualen Werte der größten Übereinstimmung zwischen Molekülen können bestimmt werden. Nach dem im Abschnitt 2.10.6 beschriebenen Verfahren wurde für verschiedene EC-Kombinationen (vgl. Tabelle 3-7) die größte gemeinsame Teilstruktur der bei den Reaktionen beteiligten Moleküle ermittelt. In Diagramm 3-20 wurden für sechs E-Werte die Ergebnisse der Untersuchungen als prozentuale Werte aufgetragen. Der Anteil der identischen Moleküle wurde bereits in Diagramm 3-19 eingetragen.

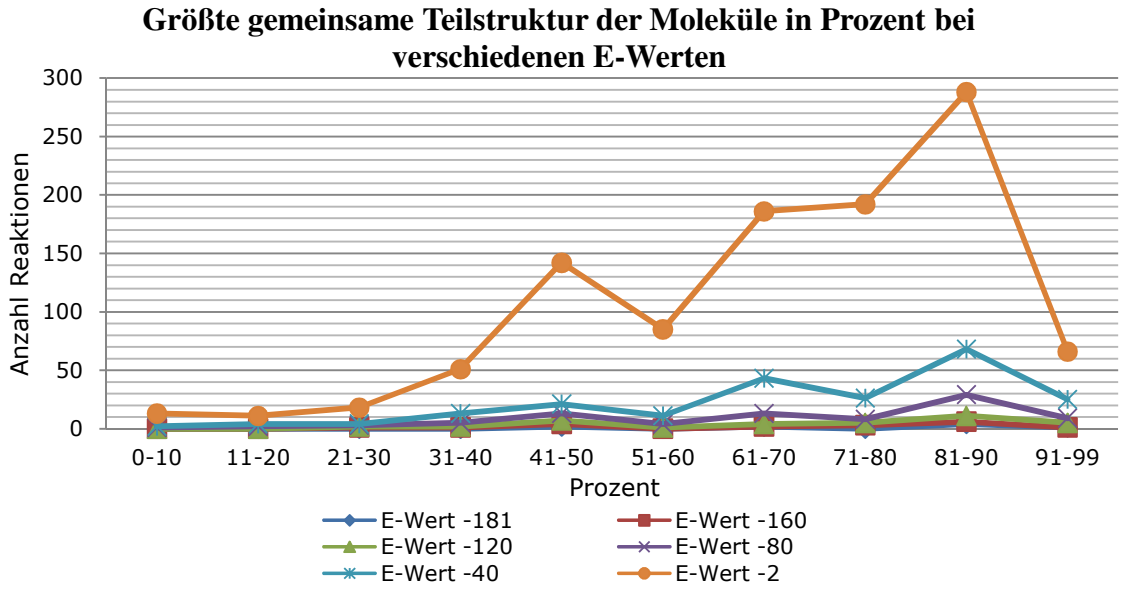


Diagramm 3-20: Ergebnis der Untersuchung der größten gemeinsamen Teilstruktur, der bei den verglichenen katalysierten Reaktionen beteiligten Moleküle. Die angegebenen E-Werten sind die Potenzen zur Basis 10.

Im Vergleich aller EC-Kombinationen fällt zunächst auf, dass der Großteil der an den Reaktionen beteiligten Moleküle identisch ist (vgl. Diagramm 3-19). Liegt ansonsten eine

prozentuale gemeinsame Teilstruktur vor, liegt diese für alle untersuchten E-Werte meist im Bereich von 81-90%. Mit ansteigendem E-Wert existieren zunehmend Moleküle, die zu einem geringeren Prozentsatz übereinstimmen. Besonders die Vergleiche bei E-Wert 10^{-2} sind auffällig, wie im Diagramm 3-20 zu erkennen ist. Dabei muss beachtet werden, dass der Anteil der Vergleiche sich mit steigendem E-Wert im Diagramm nach links verschiebt, die Kurve aber höher ist, da bei E-Wert 10^{-2} 2462 EC-Kombinationen verglichen wurden. Bei E-Wert 10^{-181} wurden die Reaktionen von 42 EC-Kombinationen verglichen. Bei E-Wert 10^{-2} stimmen bei ~290 EC-Kombinationen die beteiligten Moleküle zu maximal 90% überein. Die Anzahl hat sich bis zu einer Übereinstimmung bis maximal 50% halbiert. Bei E-Wert 10^{-40} hat sich der Wert für eine maximale Übereinstimmung von 81-90% bis maximal 50% nicht halbiert, sondern gedrittelt.

3.3.5 Einordnung der untersuchten EC-Kombinationen in die Clusterung nach identischen R-Strings

In der folgenden Tabelle 3-8 ist die Clusterung von Enzymreaktionen nach gleichen R-Strings zu sehen. Für weitere Details zu den Daten dieser Tabelle wird auf die Abschnitte 2.10.3 und 2.10.4 im Teil Daten, Algorithmen und Methoden verwiesen.

E-Wert	EC-Kombinationen	Identische R-Matrix	Anteil identischer R-Matrizen an verglichenen EC-Kombinationen in Prozent
10^{-2}	3174	127	4
10^{-40}	603	40	6,6
10^{-80}	258	20	7,8
10^{-120}	131	13	9,9
10^{-160}	71	10	14,1
10^{-181}	45	5	11,1

Tabelle 3-8: Ergebnis des Vergleichs der Clusterung von Enzymsequenzen auf Basis ähnlicher Sequenzen mit der Clusterung von Enzymen mit identischen Reaktionsmatrizen.

Bei E-Wert 10^{-2} bis E-Wert 10^{-181} wurden EC-Kombinationen, die aufgrund der Clusterung nach Sequenzähnlichkeit zustande kamen, auf die Zugehörigkeit zu Clustern identischer Reaktionsmatrizen verglichen. Bei E-Wert 10^{-181} gehören 5 von 45 EC-Kombinationen zu Clustern identischer Reaktionsmatrizen. Dies entspricht einem Anteil von 11,1%. Mit steigendem E-Wert nehmen die Anzahl der Vergleiche, sowie die Anzahl der Kombinationen

zu Clustern gleicher Reaktionsmatrizen, stetig zu. Gleichermaßen nimmt der prozentuale Anteil der EC-Kombinationen, die zu Clustern identischer Reaktionsmatrizen gehören, tendenziell ab. Der Anteil fällt von 11,1% bei E-Wert 10^{-181} bis auf 4% bei E-Wert 10^{-2} ab. Im Vergleich von E-Wert 10^{-181} mit E-Wert 10^{-160} nimmt der Anteil der EC-Kombinationen zu Clustern identischer Reaktionsmatrizen um 3% zu.

Die prozentualen Anteile der EC-Kombinationen zu Clustern gleicher Reaktionsmatrizen wurden für verschiedene E-Werte in Diagramm 3-21 aufgetragen.

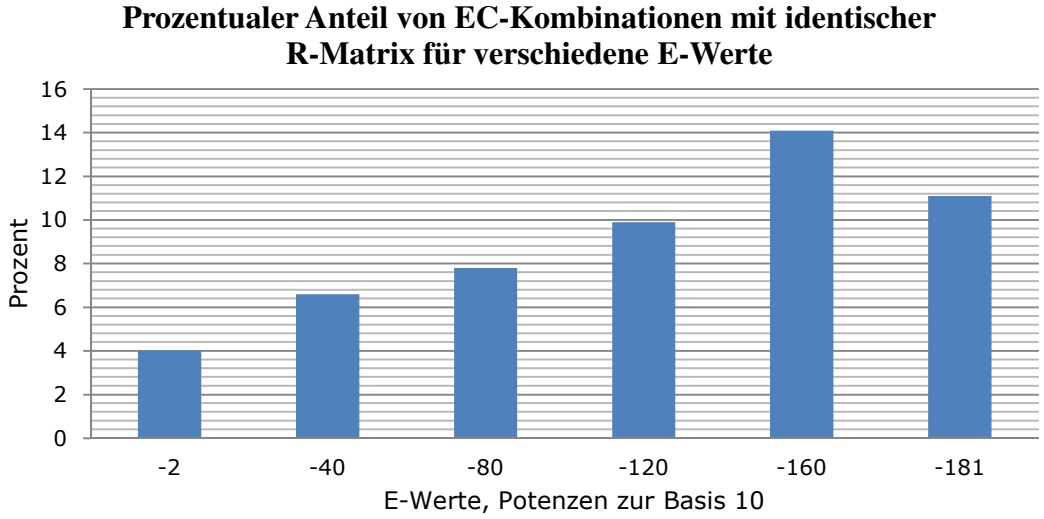


Diagramm 3-21: Ergebnis des Vergleichs von Clustern basierend auf Sequenzähnlichkeit mit Clustern basierend auf identischen Reaktionsmatrizen. Das Diagramm stellt die prozentualen Anteile der EC-Kombinationen zu Clustern identischer Reaktionsmatrizen für verschiedene E-Werte dar.

3.3.6 Übersicht über die Häufigkeit der gleicher R-Matrizen in eine Gruppennummer nach Tabelle 2-3

Für verschieden EC-Kombinationen wurde ihre Zugehörigkeit zu Clustern identischer Reaktionsmatrizen untersucht (vgl. Tabelle 2-3 und 3-8, Diagramm 3-21). Die folgende Tabelle 3-9 gibt Aufschluss, zu welcher Gruppe EC-Kombinationen gehören, falls Reaktionsmatrizen identisch sind. Die angegebene Gruppennummer bezieht sich demnach auf die Gruppennummer der Tabelle 2-3.

E-Wert	EC-Kombinationen mit identischer R-Matrix	Größte Gruppe identischer R-Matrizen	Anteil in Prozent	Gruppennummer
10^{-2}	127	47	37	2
10^{-40}	40	14	35	3
10^{-80}	20	9	45	3
10^{-120}	13	6	46,2	2
10^{-160}	10	4	40	2
10^{-181}	5	2	40	3

Tabelle 3-9: Gruppenzugehörigkeit nach Tabelle 2-3 für EC-Kombinationen mit identischer R-Matrix.

Falls EC-Kombinationen identische R-Matrizen besitzen, gehören sie am häufigsten der Gruppennummer 2 oder 3 an.

Zur Gruppe 2 gehören 17 EC Subsubklassen an. 16 dieser 17 Mitglieder gehören zur EC-Klasse 3, ein Mitglied der Gruppe gehört der EC-Klasse 2 an. Die Enzyme katalysieren die Spaltung von C-N und O-H Bindungen, sowie den Aufbau von C-O und N-H Bindungen.

Zur Gruppe 3 gehören 14 Subsubklassen an. Neun dieser 14 Mitglieder gehören zur EC-Klasse 2. Die Enzyme katalysieren die Spaltung von P-O und O-H Bindungen, sowie den Aufbau von P-O und O-H Bindungen.

3.4 Beispiele, Erklärung zu den Beispielen

In den folgenden Abschnitten werden in drei Beispielen Enzyme und deren Sequenzmuster exemplarisch für den kompletten Datensatz untersucht. Eine Übersicht des jeweiligen Clusterbaums liefert zu Beginn jedes Beispiels eine graphische Darstellung. Positionsangaben einzelner Aminosäuren beziehen sich auf die Sequenz eines Organismus‘, meist *Escherichia coli*. Aminosäuren werden mit dem „Ein-Buchstaben-Code“ abgekürzt. Die Angabe H143 bedeutet demnach Histidin an der Position 143 der Proteinsequenz. Wird eine Aminosäure nicht genauer spezifiziert, wird die Angabe „Aminosäure“ mit „AS“ abgekürzt.

In den Tabellen werden die Datenbanken TrEMBL mit „TREMBL“ und SWISS-PROT“ mit „SPROT“ abgekürzt. Stammen Sequenzen aus beiden Datenbanken, wird dies mit dem Kürzel „TRROT“ gekennzeichnet. Die Bezeichnung „Node“ bezeichnet einen Knoten des Clusterbaums. Die Angabe „c-MCS“ bezeichnet die Anzahl der gemeinsamen Atome der verglichenen Moleküle, die mit Hilfe des c-MCS Algorithmus‘ ermittelt wurden.

Bei der Vorstellung der Beispiele werden die Begriffe „konserviert“ und „hochkonserviert“ benutzt, um den Grad der Konservierung von Aminosäuren zu verdeutlichen. Diese Begriffe stehen im Zusammenhang mit den Markierungen, wie sie von CLUSTAL W in Alignments vorgenommen werden. Eine detailierte Übersicht zu den Bezeichnungen von CLUSTAL W liefert der Abschnitt 2.7.

Die erstellten Muster befinden sich in den dargestellten Clustern, sowie außerhalb der Clusterbäume farbig markiert. Dabei muss beachtet werden, dass Positionen, die im Splitcluster hochkonserviert sind, im Muttercluster möglicherweise funktionell konserviert sind. In der farbigen Markierung außerhalb der Clusterbäume wird dieser Umstand nicht beachtet. Die farbige Markierung ist eine Hilfe bei der Wiedererkennung von Aminosäuren. Im Zweifel ist immer das Muster zu beachten, dass sich innerhalb des Clusterbaums befindet. Die vollständigen Clusterbäume sind in gml-Format auf der beiliegenden DVD vorhanden.

Die dargestellten Reaktionen und Enzymnamen wurden, wie in anderen Teilen dieser Arbeit auch, der Datenbank KEGG [132] entnommen. Die verwendeten Molfiles stammen aus der Datenbank BRENDA (vgl. Abschnitt 2.11.5).

3.4.1 Beispiel 1:

Glucosamin-6-Phosphat Deaminase, EC-Nummer 3.5.99.6

Für das in diesem Beispiel vorgestellte Enzym existieren in der Datenbank 18 Sequenzmuster. Alle Sequenzmuster wurden aus Sequenzen generiert, denen die EC-Nummer 3.5.99.6 zugeordnet wurden. Das Enzym trägt die Bezeichnung Glucosamin-6-Phosphat Deaminase, EC 3.5.99.6. Die folgende Tabelle zeigt Details zu den Einträgen in der Datenbank.

Cluster-baum	Node	E-Wert	Datenbank	Ec-Nummer	Sequenz-anzahl	Muster-länge	Richtig-Positive	Falsch-Positive
21971	1	10^{-2}	TREMBL	3.5.99.6	5	273	5	0
21971	3	10^{-154}	TREMBL	3.5.99.6	2	274	2	0
24598	1	10^{-2}	TRROT	3.5.99.6	6	259	6	1
24598	3	10^{-146}	TRROT	3.5.99.6	5	261	5	1
26668	1	10^{-2}	TREMBL	3.5.99.6	4	373	4	0
19905	634	10^{-23}	TRROT	3.5.99.6	9	107	9	0
19905	748	10^{-33}	TREMBL	3.5.99.6	2	92	4	0
19905	844	10^{-40}	TREMBL	3.5.99.6	3	132	3	0
19905	935	10^{-46}	TRROT	3.5.99.6	8	116	8	0
19905	974	10^{-51}	TRROT	3.5.99.6	4	125	4	0
19905	987	10^{-51}	TRROT	3.5.99.6	11	81	11	0
19905	1021	10^{-54}	TREMBL	3.5.99.6	2	158	2	0
19905	1036	10^{-55}	TRROT	3.5.99.6	9	148	9	0
19905	1063	10^{-57}	TREMBL	3.5.99.6	5	145	5	0
19905	1081	10^{-59}	TRROT	3.5.99.6	82	40	82	34
19905	1203	10^{-85}	TREMBL	3.5.99.6	2	243	2	0
19905	1257	10^{-94}	TRROT	3.5.99.6	39	131	39	20
19905	1432	10^{-144}	SPROT	3.5.99.6	4	286	4	0

Tabelle 3-10: Sequenzmuster für EC-Nummer 3.5.99.6.

Die 18 Sequenzmuster für ausschließlich EC-Nummer 3.5.99.6 befinden sich in den vier Clusterbäumen 21971, 24598, 26668 und 19905. Es existieren jeweils zwei Muster in den Clusterbäumen 21971 und 24598, ein Muster im Clusterbaum 26668 und 13 Muster im Clusterbaum 19905. Im Clusterbaum 21971 sowie in den Clusterbäumen 24598 und 26668 befinden sich ausschließlich Sequenzen, die die EC-Nummer 3.5.99.6 tragen. Im Clusterbaum 19905 wurden zusätzlich, bei verschiedenen E-Werten, die EC-Nummern 1.1.1.49, EC 3.1.1.17, EC 3.1.1.31 und EC 5.3.1.10 geclustert. Im Clusterbaum 19905 befinden sich damit Enzyme aus den EC-Klassen der Oxidoreduktasen, der Hydrolasen und der Isomerasen. Von allen Enzymsequenzen eines Clusters, in denen sich diese EC-Nummern befinden, wurde kein Muster erstellt, da die Anzahl der Sequenzen über 1000 liegt.

In den folgenden Ausführungen wird besonders auf die Muster des Clusterbaums 19905 eingegangen. Der Clusterbaum besteht aus 1494 Knoten. In den farbig markierten Clustern sind mehr als eine EC-Nummer enthalten.

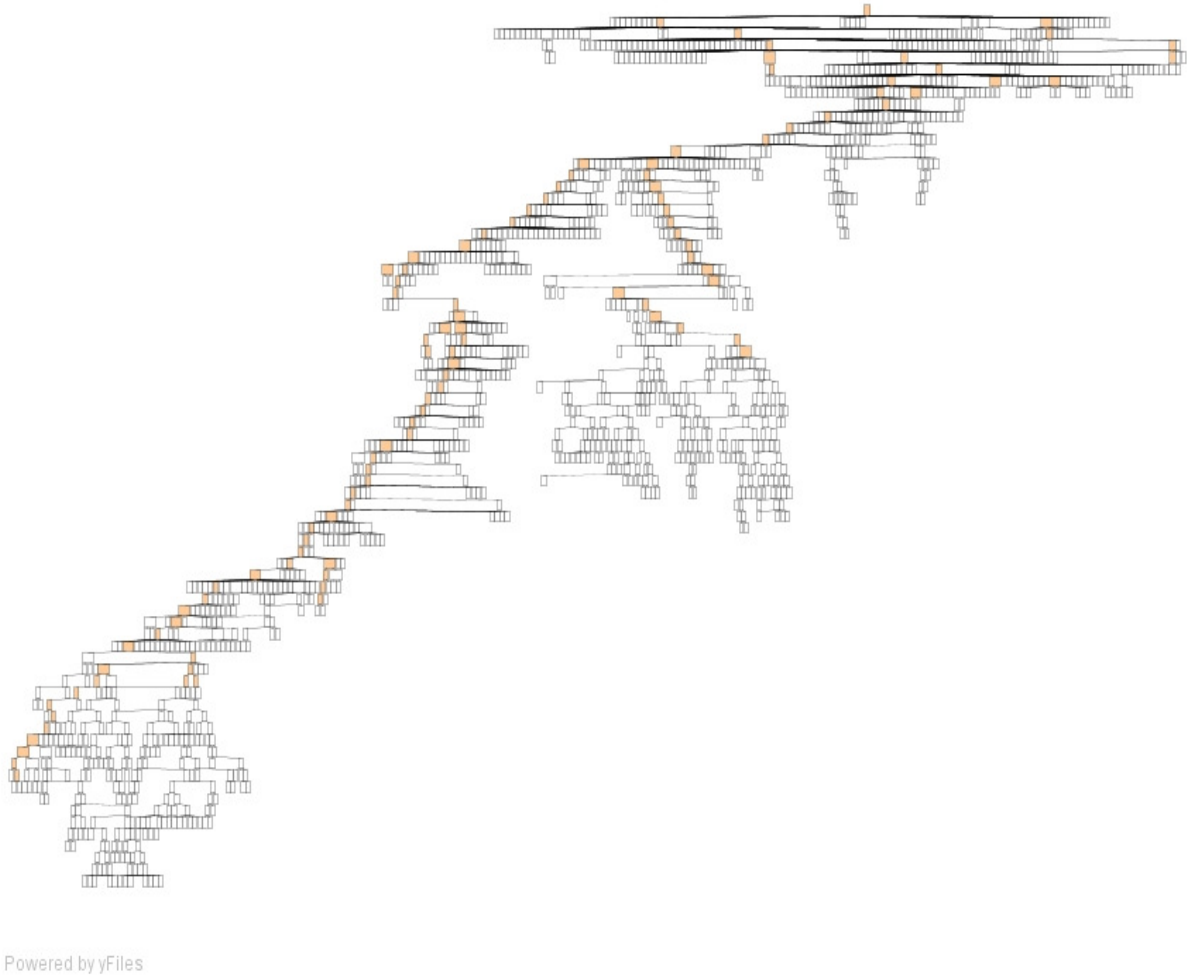


Abbildung 3-1: Der Clusterbaum 19905**.**

Aus einem Cluster des Clusterbaums 19905 wurde ein Muster für die EC-Nummern 3.5.99.6 und 1.1.1.49 erstellt.

Cluster-baum	Node	E-Wert	Datenbank	Ec-Nummer	Sequenz-anzahl	Muster-länge	Richtig-Positive	Falsch-Positive
19905	304	10^{-9}	TREMBL	3.5.99.6 1.1.1.19	2	112	2	1

Tabelle 3-11: Muster für EC 3.5.99.6 und EC 1.1.1.49 im Clusterbaum 19905.

Aus zwei Clustern des Clusterbaums 19905 wurden zwei Muster für die EC-Nummern 3.5.99.6 und 3.1.1.31 erstellt.

Cluster-baum	Node	E-Wert	Datenbank	Ec-Nummer	Sequenz anzahl	Muster-länge	Richtig-Positive	Falsch-Positive
19905	269	10^{-8}	TRROT	3.5.99.6 3.1.1.31	8	65	8	6
19905	697	10^{-29}	TREMBL	3.5.99.6 3.1.1.31	4	94	4	1

Tabelle 3-12: Muster für EC 3.5.99.6 und EC 3.1.1.31 im Clusterbaum 19905.

Aus 13 Clustern dieses Clusterbaums wurden Muster für die EC-Nummern 3.5.99.6 und 5.3.1.10 erstellt.

Cluster-Baum	Node	E-Wert	Datenbank	Ec-Nummer	Sequenz-anzahl	Muster-länge	Richtig-Positive	Falsch-Positive
19905	638	10^{-24}	TREMBL	3.5.99.6 5.3.1.10	2	106	2	2
19905	697	10^{-29}	TREMBL	3.5.99.6 5.3.1.10	4	94	4	1
19905	698	10^{-29}	TRROT	3.5.99.6 5.3.1.10	155	15	155	150
19905	737	10^{-32}	TRROT	3.5.99.6 5.3.1.10	154	16	154	144
19905	827	10^{-39}	TRROT	3.5.99.6 5.3.1.10	154	16	154	144
19905	910	10^{-45}	TRROT	3.5.99.6 5.3.1.10	151	16	154	144
19905	951	10^{-49}	TRROT	3.5.99.6 5.3.1.10	6	176	6	0
19905	1005	10^{-53}	TRROT	3.5.99.6 5.3.1.10	141	29	141	129
19905	1025	10^{-54}	TRROT	3.5.99.6 5.3.1.10	141	29	141	129
19905	1043	10^{-56}	TRROT	3.5.99.6 5.3.1.10	141	29	141	129
19905	1068	10^{-58}	TRROT	3.5.99.6 5.3.1.10	130	33	135	122
19905	1247	10^{-93}	TRROT	3.5.99.6 5.3.1.10	46	122	46	29
19905	1255	10^{-94}	TRROT	3.5.99.6 5.3.1.10	7	261	7	2

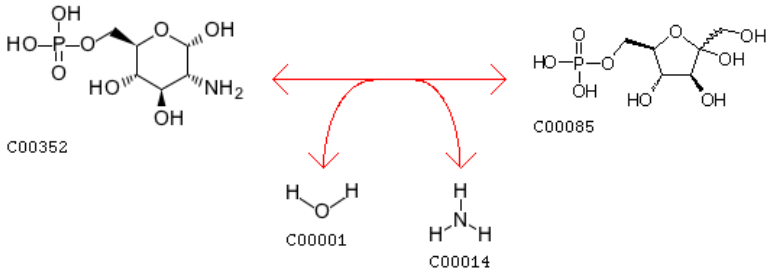
Tabelle 3-13: Muster für EC 3.5.99.6 und EC 5.3.1.10 im Clusterbaum 19905.

3.4.1.1 Eigenschaften von EC 3.5.99.6

Name des Enzyms: Glucosamin-6-Phosphat Deaminase

Das Enzym Glucosamin-6-Phosphat Deaminase, EC-Nummer 3.5.99.6, katalysiert die Desaminierung von D-Glucosamin-6-Phosphat unter Verwendung von Wasser. Dabei entstehen die Moleküle D-Fructose-6-Phosphat und Ammoniak.

Reaktion: D-Glucosamin-6-Phosphat + H_2O ↔ D-Fructose-6-Phosphat + NH_3



Die vollständige Sequenz von Glucosamin-6-Phosphat Deaminase variiert je nach Organismus zwischen 230 und 270 Aminosäuren und erreicht in *Rhodopirellula baltica* eine Länge von 637 Aminosäuren. Es kann als Monomer, so in *Candida albicans* und *Giardia intestinalis* mit einem Molekulargewicht von 17,5 kDa bzw. 29 kDa, oder als Hexamer, zum Beispiel in *Escherichia coli* mit einem Molekulargewicht von 29,7 kDa pro Untereinheit, vorkommen [140]. In *E. coli* bildet das Enzym ein Dimer aus zwei Trimeren [144]. Glucosamin-6-Phosphat Deaminase aus *E. coli* ist ein allosterisches Enzym und gehört nach dem Modell von Monod-Wayman-Changeux zum K-Typ [145] und wird von N-Acetylglucosamin-6-Phosphat aktiviert [146]. Nach diesem Modell existieren sowohl ein T-Zustand (T=tense) als auch ein R-Zustand (R=relaxed) des Enzyms. Das Substrat bindet ausschließlich im R-Zustand [147].

3.4.1.2 Der katalytische Mechanismus von EC 3.5.99.6

Nach Studien von Midelfort und Rose [148] wurde ein Mechanismus postuliert, der nach Arbeiten von Oliva *et al.* [146] in strukturellen Kontext gebracht wurde. Da es sich um ein allosterisches Enzym handelt, wurde neben dem aktivem Zentrum in weiteren Arbeiten auch das allosterische Zentrum des Enzyms untersucht. Es zeigte sich, wie im Muster der PROSITE-Datenbank angedeutet wird, dass H143 eine entscheidende Rolle bei der Ringöffnung des Pyranosemoleküls spielt. Darüber hinaus steht im R-Zustand die Imidazolgruppe des Histidins in Kontakt zu den Carboxy-Gruppen von D141 und E148 [146]. Dies führt zur Umformung des Enzyms zum T-Zustand, da sich durch die Bindung zum Histidin die Seitenkette von E148 in Richtung des allosterischen Zentrums bewegt und eine Salzbrücke mit K160 eingeht. In Folge dieser Bewegung verliert E148 den Kontakt zum Histidin und eine Verschiebung des Loops der Aminosäuren 144-154 findet statt [149]. Die Aminosäuren D141, H143 und E148 spielen somit eine Schlüsselrolle bei der katalytischen Reaktion. Die Bedeutung der Aminosäuren wurde in Mutationsexperimenten untermauert [150]. Im aktiven Zentrum sind zudem die Aminosäuren K208, R172, Y85 und D72 vorhanden [149]. Eine besondere Rolle spielen die Aminosäuren S151, R158 und K160, die sich im allosterischen Zentrum befinden und dort beteiligt sind, die Phosphatgruppe des Aktivators N-Acetylglucosamin-6-Phosphat zu binden [151]. Die Aminosäuren Y121 und Y254 sind für den Übergang vom T- zum R-Zustand wichtig, was in Mutationsversuchen gezeigt wurde [152]. Die Abbildung 3-2 zeigt, farbig hervorgehoben, die wichtigen Aminosäuren des allosterischen und des aktiven Zentrums.

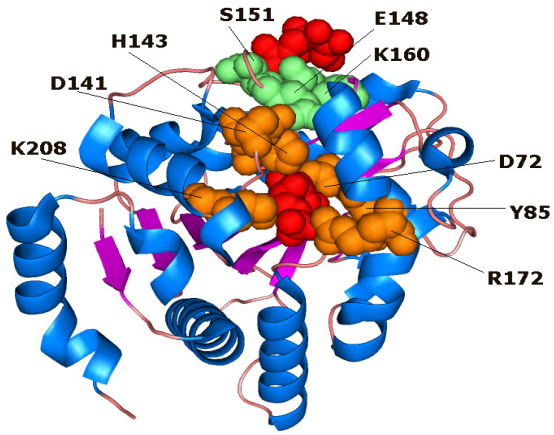


Abbildung 3-2: Glucosamin-6-Phosphat Deaminase, PDB 1FQO, Monomer, *E. coli*. Wichtige Aminosäuren des aktiven Zentrums grün, wichtige Aminosäuren des allosterischen Zentrums orange, α-Helixes blau, β-Faltblätter pink, Fructose-6-Phosphat (rot) bindet an beide Zentren. Die Aminosäure R158 ist verdeckt und wurde nicht beschriftet.

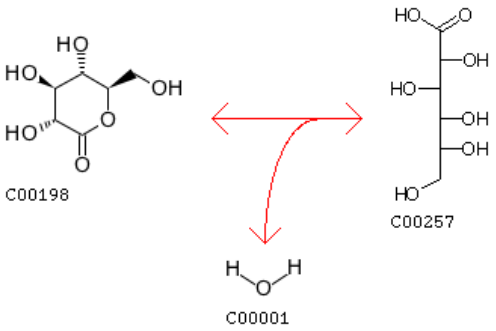
3.4.1.3 Weitere EC-Nummern im Clusterbaum und deren katalysierte Reaktionen

EC-Nummer 3.1.1.17:

Name des Enzyms: D-Glucono-1,5-Lacton Lactonohydrolase

Das Enzym D-Glucono-1,5-Lacton Lactonohydrolase katalysiert die Hydrolyse von D-Glucono-1,5-Lacton. Dabei entsteht das Molekül D-Gluconat.

Reaktion: D-Glucono-1,5-Lacton + H_2O ↔ D-Gluconat

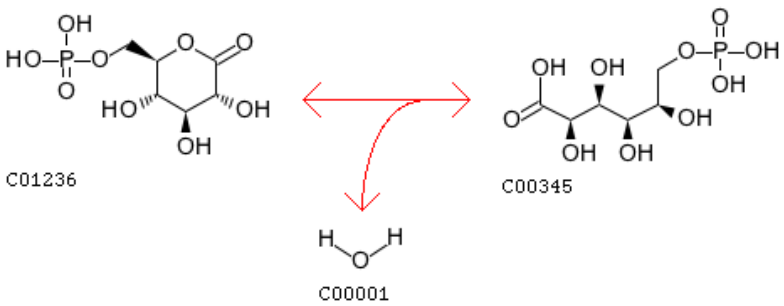


EC-Nummer 3.1.1.31:

Name des Enzyms: 6-Phospho-D-Glucono-1,5-Lacton Lactonohydrolase

Das Enzym 6-Phospho-D-Glucono-1,5-Lacton Lactonhydrolase katalysiert die Hydrolyse von D-Glucono-1,5-Lacton 6-Phosphat. Dabei entsteht das Molekül 6-Phospho-D-Gluconat.

Reaktion: D-Glucono-1,5-Lacton 6-Phosphat + H_2O ↔ 6-Phospho-D-Gluconat

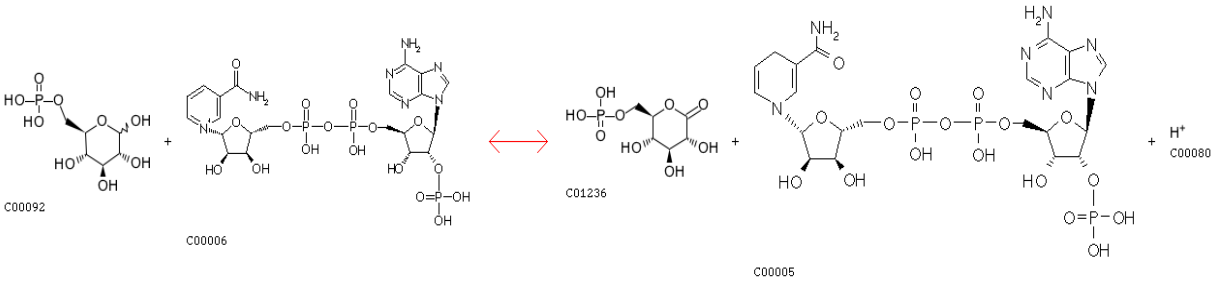


EC-Nummer 1.1.1.49:

Name des Enzyms: Glucose-6-Phosphat 1-Dehydrogenase

Das Enzym Glucose-6-Phosphat 1-Dehydrogenase katalysiert die Oxidation von Glucose-6-Phosphat. $NADP^+$ agiert als Wasserstoff-Akzeptor.

Reaktion: D-Glucose-6-Phosphat + $NADP^+$ ↔ D-Glucono-1,5-Lacton 6-Phosphat + NADPH + H^+



EC-Nummer 5.3.1.10:

Die EC-Nummer 5.3.1.10 wurde im Jahr 1961 erstellt und im Jahr 2000 gelöscht. Die EC-Nummer entspricht nun der EC-Nummer 3.5.99.6. Obwohl die EC-Nummer seit einigen Jahren nicht mehr existiert, ist sie im Clusterbaum vorhanden, da TrEMBL die Angaben nicht aktualisiert hat.

Reaktion:

Da die EC-Nummer 5.3.1.10 zu EC-Nummer 3.5.99.6 transferiert wurde, entspricht die Reaktion von EC 5.3.1.10 der katalysierten Reaktion von EC 3.5.99.6.

3.4.1.4 Größte gemeinsame Teilstrukturen der Edukte, Produkte und Co-Substrate der Enzymreaktionen

Die größten gemeinsamen Teilstrukturen der Edukte, Produkte und Co-Substrate werden für die Reaktionen der EC-Nummern EC 3.5.99.6 + EC 3.1.1.31, EC 3.5.99.6 + EC 1.1.1.49 und EC 3.1.1.31 + EC 1.1.1.49 ermittelt. Die Ergebnisse der Untersuchungen werden in den folgenden Tabellen zusammengefasst. Die Darstellungen beschränken sich auf Vergleiche, bei denen die größten gemeinsamen Teilstrukturen ermittelt wurden.

Vergleich EC 3.5.99.6 EC 3.1.1.31	**Molfile**	**Atome**	**Molekül**
	9143.mol	16	D-Fructose 6-Phosphat
	5754.mol	16	6-Phospho-D-Glucono-1,5-Lacton
c-MCS		16	
	9198.mol	16	D-Glucosamin 6-Phosphat
	5753.mol	17	6-Phospho-D-Gluconat
c-MCS		15	
	9143.mol	16	D-Fructose 6-Phosphat
	5753.mol	17	6-Phospho-D-Gluconat
c-MCS		16	

Tabelle 3-14: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 3.5.99.6 und EC 3.1.1.31.

Vergleich EC 3.5.99.6 EC 1.1.1.49	**Molfile**	**Atome**	**Molekül**
	9143.mol	16	D-Fructose 6-Phosphat
	9191.mol	16	D-Glucono-1,5-Lacton 6-Phosphat
c-MCS		16	
	9143.mol	16	D-Fructose 6-Phosphat
	9206.mol	16	D-Glucose 6-Phosphat
c-MCS		16	
	9198.mol	16	D-Glucosamin 6-Phosphat
	9191.mol	16	D-Glucono-1,5-Lacton 6-Phosphat
c-MCS		15	

Tabelle 3-15: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 3.5.99.6 und EC 1.1.1.49.

Vergleich EC 3.1.1.31 EC 1.1.1.49	**Molfile**	**Atome**	**Molekül**
	5754.mol	16	6-Phospho-D-Glucono-1,5-Lacton
	9206.mol	16	D-Glucose 6-Phosphat
c-MCS		16	
	5754.mol	16	6-Phospho-D-Glucono-1,5-Lacton
	9191.mol	16	D-Glucono-1,5-Lacton 6-Phosphat
c-MCS		16	
	5753.mol	17	6-Phospho-D-Gluconat
	9191.mol	16	D-Glucono-1,5-Lacton 6-Phosphat
c-MCS		16	

Tabelle 3-16: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 3.1.1.31 und EC 1.1.1.49.

Wie den Tabellen 3-14 bis 3-16 entnommen werden kann, sind die Edukte, Produkte oder Co-Substrate der untersuchten Enzymreaktionen identisch oder sehr ähnlich. Mindestens ein Molekül ist bei jedem Vergleich identisch.

3.4.1.5 Vergleich der EC-Kombinationen mit der Clusterung nach identischen R-Strings

Die Enzyme mit den EC Nummern EC 3.5.99.6, EC 1.1.1.49 und EC 3.1.1.31 wurden aufgrund identischer R-Strings nicht gruppiert.

3.4.1.6 Untersuchung von Falsch-Positiven Treffern ausgewählter Sequenzmuster

Überblick über die Treffer des Musters EC 3.5.99.6, E-Wert 10^{-59}, Clusterbaum 19905. Bei der Suche gegen SWISS-PROT wurden keine Falsch-Positiven Treffer detektiert.

Sequenz	Start	Ende	Datenbank	Bezeichnung
A0IZR1_9GAMM	66	234	TREMBL	Glucosamine-6-phosphate isomerase
A0JY49_ARTS2	66	234	TREMBL	Glucosamine-6-phosphate isomerase
A0RI59_BACAH	62	227	TREMBL	Glucosamine-6-phosphate isomerase
Q033M5_LACC3	61	226	TREMBL	Glucosamine-6-phosphate isomerase
Q2B733_9BACI	67	235	TREMBL	N-acetylglucosamine-6-phosphate isomerase
Q1YAR0_STAAU	66	235	TREMBL	Glucosamine/Galactosamine-6-Phoshat isomerase
Q2BT67_LACRE	16	181	TREMBL	Glucosamine-6-phosphate isomerase
Q38KB9_LACRE	61	226	TREMBL	6-Phosphogluconolactonase
Q62N85_BACLD	66	234	TREMBL	N-acetylglucosamine-6-phosphate isomerase
Q2AH48_9FIRM	84	252	TREMBL	Glucosamine-6-phosphate isomerase
Q2ARQ2_9BACI	62	227	TREMBL	Glucosamine-6-phosphate isomerase
Q2WRZ1_CLOBE	66	234	TREMBL	Glucosamine-6-phosphate isomerase
Q301B5_STRSU	61	226	TREMBL	Glucosamine/Galactosamine-6-Phoshat isomerase
Q3CA99_9CLOT	66	234	TREMBL	Glucosamine-6-phosphate isomerase
A0WNC8_9GAMM	66	237	TREMBL	Glucosamine-6-phosphate isomerase
A0X805_9GAMM	77	246	TREMBL	Glucosamine-6-phosphate isomerase
Q02Y08_LACLS	61	226	TREMBL	Glucosamine-6-phosphate isomerase
Q03LS1_STRTD	61	226	TREMBL	Glucosamine-6-phosphate isomerase
Q040I7_LACGA	69	234	TREMBL	Glucosamine-6-phosphate isomerase
Q0ERD7_THEET	66	234	TREMBL	Glucosamine-6-phosphate isomerase
Q2G0K8_STAA8	66	235	TREMBL	Hypothetical protein
Q1Y359_STAAU	66	235	TREMBL	Glucosamine/Galactosamine-6-Phoshat isomerase
Q33TC0_9GAMM	66	234	TREMBL	Glucosamine-6-phosphate isomerase
Q2E5T0_BACCE	62	227	TREMBL	Glucosamine-6-phosphate isomerase
A0WZK0_9GAMM	66	267	TREMBL	Glucosamine-6-phosphate isomerase
Q047I3_LACDB	61	226	TREMBL	Glucosamine-6-phosphate isomerase
Q1UAW3_LACRE	61	226	TREMBL	Glucosamine/Galactosamine-6-Phoshat isomerase
Q3CGD4_THEET	66	234	TREMBL	Glucosamine-6-phosphate isomerase
Q3Y0E9_ENTFC	61	226	TREMBL	Glucosamine-6-phosphate isomerase
A2RDX6_STRP5	61	226	TREMBL	Glucosamine-6-phosphate isomerase
Q74HC4_LACJO	61	226	TREMBL	Glucosamine-6-phosphate isomerase
Q5M5E7_STRT2	86	251	TREMBL	Glucosamine-6-phosphate isomerase
Q6NJ91_CORDI	64	233	TREMBL	Putative Isomerase
Q5M0W0_STRT1	86	251	TREMBL	Glucosamine-6-phosphate isomerase

Tabelle 3-17: Falsch-Positive Treffer für EC 3.5.99.6 aus Clusterbaum 19905, E-Wert 10^{-59}.

Überblick über die Treffer des Musters für EC 3.5.99.6, E-Wert 10^{-94}, Clusterbaum 19905.

Sequenz	Start	Ende	Datenbank	Bezeichnung
Q0SP13_BORAP	1	240	TREMBL	Glucosamine-6-phosphate isomerase
Q0T6S6_SHIF8	1	240	TREMBL	Glucosamine-6-phosphate deaminase
Q1CKN7_YERPN	1	240	TREMBL	Glucosamine-6-phosphate isomerase
A0W3N0_9ENTR	1	240	TREMBL	Glucosamine-6-phosphate isomerase
A1JQE8_YERE8	1	240	TREMBL	Putative Glucosamine-6-phosphate isomerase
Q1C537_YERPA	1	240	TREMBL	Putative Glucosamine-6-phosphate isomerase
A2UJR8_ECOLI	1	240	TREMBL	Glucosamine-6-phosphate isomerase
Q324M6_SHIBS	1	240	TREMBL	Glucosamine-6-phosphate deaminase
Q57RQ0_SALCH	1	240	TREMBL	Glucosamine-6-phosphate deaminase
Q6LTD4_PHOPR	1	240	TREMBL	Putative Glucosamine-6-phosphate isomerase
Q6LKE7_PHOPR	1	240	TREMBL	Putative Glucosamine-6-phosphate isomerase
Q0WDQ5_YERPE	1	240	TREMBL	Putative Glucosamine-6-phosphate isomerase
Q66DC7_YERPS	1	240	TREMBL	Putative Glucosamine-6-phosphate isomerase
Q6LKE7_PHOPR	1	240	TREMBL	Putative Glucosamine-6-phosphate isomerase
Q3Z4C2_SHISS	1	240	TREMBL	Glucosamine-6-phosphate deaminase
Q64XP2_BACFR	1	240	TREMBL	Glucosamine-6-phosphate isomerase
A1A8T7_ECOK1	1	240	TREMBL	Glucosamine-6-phosphate deaminase
Q32IQ2_SHIDS	1	240	TREMBL	Glucosamine-6-phosphate deaminase
Q5PCH6_SALPA	1	240	TREMBL	Glucosamine-6-phosphate isomerase
Q662L3_BORGA	1	240	TREMBL	Glucosamine-6-phosphate isomerase

Tabelle 3-18: Falsch-Positive Treffer für EC 3.5.99.6 aus Clusterbaum 19905, E-Wert 10^{-94}.

Aus den Bezeichnungen der getroffenen Sequenzen (vgl. Tabellen 3-17 und 3-18) kann geschlossen werden, dass zumeist Sequenzen des Enzyms Glucosamin-6-Phosphat Deaminase/Isomerase getroffen werden. Diese wurde als Falsch-Positiv bewertet, da in der Datenbank TrEMBL keine EC-Nummern angegeben wurden. Eine Übersicht der Kriterien, die für Beurteilung von Treffern maßgebend sind, liefert der Abschnitt 2.9 im Teil Daten, Algorithmen und Methoden.

3.4.1.7 PROSITE- und Clustermuster

Für das Enzym EC Glucosamin-6-Phosphat Deaminase, EC 3.5.99.6, ist der PROSITE-Eintrag PS01161, PROSITE Dokumentation PDOC00894, vorhanden. Das dort vorgeschlagene Muster stammt aus einem konservierten Bereich aus der Mitte der Sequenz. Dabei wird besonders auf die Aminosäure Histidin hingewiesen, die während der Katalyse eine wichtige Funktion bei der Öffnung des Pyranoserings übernimmt. Diese Aminosäure befindet sich an letzter Stelle des elfstelligen Musters der PROSITE-Datenbank.

```
[LIVM]-x(3)-[GNH]-x(0,1)-[LITCRV]-x-[LIVWF]-x-[LIVMF]-x-[GS]-[LIVM]-G-x-
[DENV]-G-[HN]
```

Abbildung 3-3: PROSITE-Muster für EC 3.5.99.6, PS01161.

Die Wissenschaftler der PROSITE-Datenbank haben mit diesem Muster nach Treffern in der Sequenzdatenbank SWISS-PROT gesucht. Daraus resultieren 161 Richtig-Positive und 9 Falsch-Positive Treffer. Falsch-Positive Treffer sind nach Definition von PROSITE fehlende Treffer des Musters auf Sequenzen, die die EC-Nummer des Musters tragen.

Im Clusterbaum befinden sich die Sequenzen verschiedener Organismen. Die aus dem Clusterbaum 19905 näher untersuchten Muster für EC 3.5.99.6 (vgl. Tabelle 3-17 und 3-18) sind wesentlich länger als das Muster der PROSITE-Datenbank. Beide Muster werden in den Abbildungen 3-4 und 3-5 dargestellt.

Clustermuster für EC 3.5.99.6, E-Wert 10^{-59}, Clusterbaum 19905, 40 Musterpositionen.

```
L-D-E-Y-x(2)-[ILV]-x(7)-Y-x(3)-[ILM]-x(28,40)-Y-[DEN]-x(2)-[ILV]-x(7,8)-Q-
[ILV]-L-G-[ILV]-G-x-[DN]-x-H-[IV]-x-F-N-E-P-x(7)-T-x(4)-L-x(3)-T-x(3)-N-
x(3)-F-x(6,7)-P-x(6)-G-x(3)-[IL]-x(5)-[ILMV]-[ILMV]-[LMV]-x(3)-G-x(2)-K-
x(18)-[ST]-x-[LM]-x(2)-H-x(7)-D
```

Abbildung 3-4: Muster für EC 3.5.99.6, E-Wert 10^{-59}. Positionen, die auch im PROSITE-Muster vorhanden sind, wurden rot markiert.

Clustermuster für EC 3.5.99.6, E-Wert 10^{-94}, Clusterbaum 19905, 131 Musterpositionen.

```
M-[KR]-x-[IV]-x(9,10)-W-x-[AS]-x-[HY]-[IV]-x(3)-I-x(4)-P-[ST]-x(4)-F-[ILV]-
L-G-x-P-T-G-x-[ST]-P-[ILM]-x(2)-Y-[EKQ]-x-L-[IV]-x(3)-[EKNQ]-x(3)-[ILV]-S-F-
[EKQ]-x-V-[IV]-x-F-x-M-D-E-Y-[IV]-x-[ILM]-x(6)-S-Y-x(2)-[FY]-[LM]-x(2)-N-F-
[FI]-x-[HK]-[IV]-[DN]-I-x(2)-[EKQ]-N-x-[HN]-[FIL]-L-[DN]-G-x(3)-[DN]-x(4)-C-
x(2)-Y-E-x(2)-I-x(3)-G-x-I-x-L-F-[ILMV]-G-G-[IV]-G-x-D-G-H-[IV]-A-F-N-E-P-x-
S-S-[FL]-x-S-R-T-R-x-K-x-L-[ST]-x-[DE]-T-x(3)-N-[AS]-R-F-F-x(6)-V-P-[EKQ]-x-
[AS]-L-T-[IV]-G-[ILV]-x-T-[ILV]-[LM]-x-[AS]-x-[EK]-[IV]-[ILM]-[IL]-[ILM]-
x(2)-G-x(2)-K-x(2)-A-[LV]-x(3)-[IV]-E-x(2)-[IV]-x-[HQ]-x-W-x-[IV]-[ST]-x-L-
Q-[FLM]-H-x(5)-[IV]-x-D
```

Abbildung 3-5: Muster für EC 3.5.99.6, E-Wert 10^{-94}. Positionen, die auch im PROSITE-Muster vorhanden sind, wurden rot markiert.

3.4.1.8 Identifizierung des PROSITE-Musters in der 3D-Darstellung des Enzyms

Die folgende Abbildung zeigt die 3D-Struktur des Enzyms Glucosamin-6-Phosphat Deaminase. Die Positionen des PROSITE-Musters wurden farbig markiert.

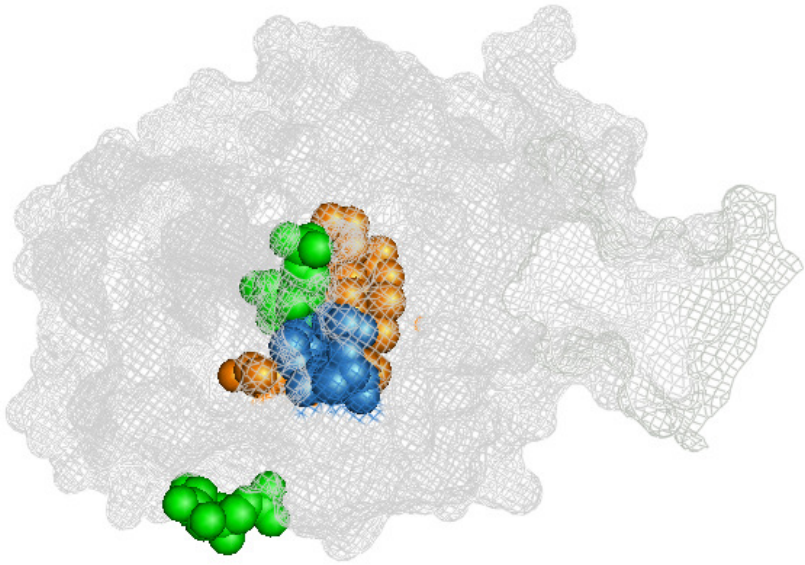


Abbildung 3-6: Glucosamin-6-Phosphat Deaminase, PDB 1FQO, Monomer, *E. coli*. Hochkonservierte Positionen wurden orange, konservierte Positionen blau, Fructose-6-Phosphat wurde grün markiert. Das Gerüst des übrigen Moleküls ist grau hinterlegt.

In der dargestellten Abbildung wurden die Positionen des PROSITE-Musters farbig markiert. Zusätzlich sind zwei Moleküle Fructose-6-Phosphat zu sehen. Das PROSITE-Muster liegt in der Mitte der Sequenz im Sequenzbereich I125 bis H143 und markiert eine Auswahl der konservierten Positionen des aktiven Zentrums. In diesem Bereich befinden sich alle in der Literatur beschriebene, für die katalytische Reaktion wichtige, Aminosäuren. Histidin an Position 143 befindet sich an letzter Stelle des Musters, allerdings gibt das Muster für diese Position an, dass hier auch die Aminosäure Asparagin stehen kann. Vor diesem Bereich befindet sich die Aminosäure Y121, die für den Übergang des Moleküls vom T- zum R-Zustand wichtig [152] und nicht im Muster enthalten ist. Die Aminosäuren D72 und Y85 des aktiven Zentrums sind ebenfalls nicht vorhanden. Nach diesem Bereich befinden sich die katalytisch wichtigen Aminosäuren E148, S151, R158, K160, R172 und K208, die bei der Bildung des aktiven Zentrums beteiligt sind und ebenfalls nicht für das PROSITE-Muster ausgewählt wurden.

3.4.1.9 Identifizierung des generierten Musters, E-Wert 10^{-94}, Clusterbaum 19905, in der 3D-Darstellung des Enzyms

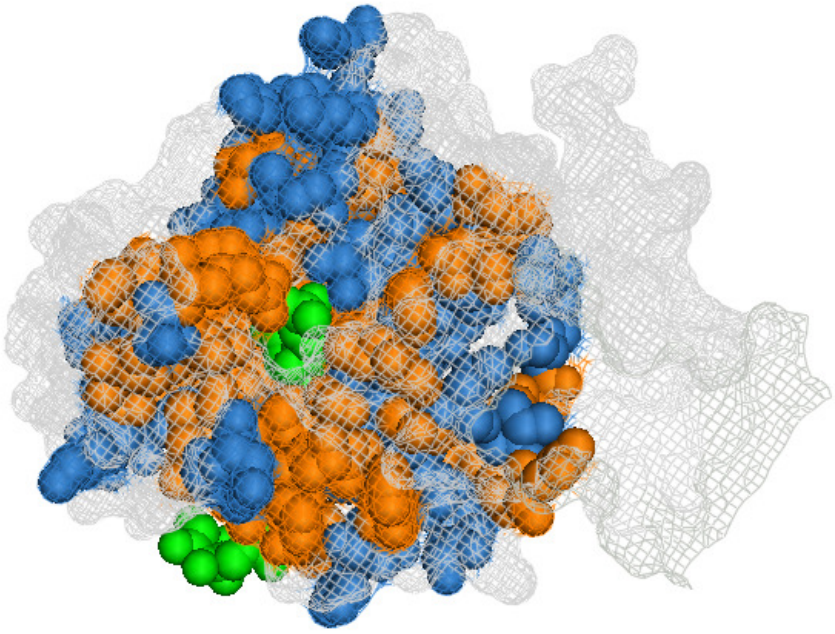


Abbildung 3-7: Glucosamin-6-Phosphat Deaminase, PDB 1FQO, Monomer, *E. coli*. Muster aus Clusterbaum 19905, E-Wert 10^{-94}. Konservierte Musterpositionen wurden blau, hochkonservierte Positionen wurden orange und Fructose-6-Phosphat wurde grün markiert.

Die Positionen des generierten Musters für das Enzym Glucosamin-6-Phosphat Deaminase, EC 3.5.99.6, wurden in der 3D-Struktur PDB 1FQO, hervorgehoben. Das dargestellte Muster wurde aus 39 Sequenzen generiert, die bei E-Wert 10^{-94} geclustert wurden. Das Muster besteht aus 131 Positionen und ist vor dem Hintergrund, dass die vollständige Sequenz in *E.coli* aus 266 Aminosäuren besteht, relativ lang. Das Muster deckt den Sequenzbereich AS 1 bis AS 240 ab und erstreckt sich damit annähernd über die vollständige Sequenz. Innerhalb dieses Musters ist der Bereich G129 bis F174 besonders stark konserviert. 35 von 131 Positionen des Musters befinden sich in diesem Bereich. Dies ist exakt der Bereich des aktiven Zentrums des Enzyms. In der Abbildung 3-7 ist die Konzentrierung der Musterpositionen auf den Bereich des aktiven Zentrums gut zu erkennen. Alle Aminosäuren, die nach Literaturrecherche wichtig für die Funktion des Enzyms sind (vgl. Abschnitt 3.4.1.2), sind in diesem Muster markiert. Dazu gehören auch Aminosäuren, die im Bereich des allosterischen Zentrums liegen. Im Randbereich des Moleküls befinden sich keine markierten, konservierten Positionen.

3.4.1.10 Identifizierung des generierten Musters, E-Wert 10^{-59}, Clusterbaum 19905, in der 3D-Darstellung des Enzyms

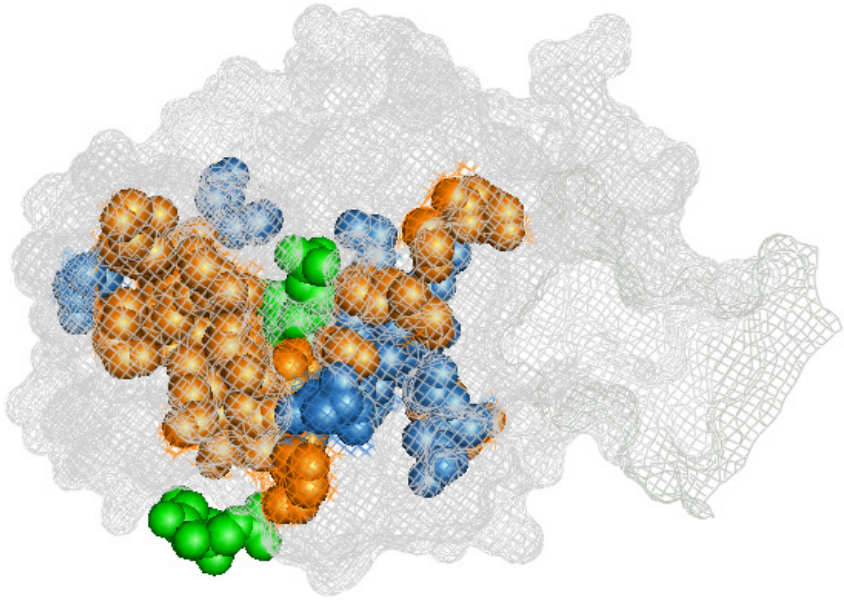


Abbildung 3-8: Glucosamin-6-Phosphat Deaminase, PDB 1FQO, Monomer, *E. coli*. Muster aus Clusterbaum 19905, E-Wert 10^{-59}. Konservierte Musterpositionen wurden blau, hochkonservierte Positionen wurden orange und Fructose-6-Phosphat wurde grün markiert.

Das in der Abbildung 3-8 hervorgehobene Muster für das Enzym Glucosamin-6-Phosphat Deaminase ist im Vergleich zu dem entsprechenden PROSITE-Muster länger, aber im Vergleich zu dem Muster, das aus Sequenzen des gleichen Clusterbaums bei E-Wert 10^{-94} entstanden ist, um etwa ein Drittel kürzer. Das Muster, das aus 40 Positionen besteht, erstreckt sich auf der Sequenz von M71 bis D240. Wie schon das Muster bei E-Wert 10^{-94}, konzentrieren sich die Musterpositionen auf den Bereich des aktiven Zentrums, wie anhand der Abbildung 3-8 zu sehen ist. Dabei ist eine noch stärkere Konzentrierung der Musterpositionen auf diesen Bereich zu erkennen, als schon das generierte Muster bei 10^{-94} zeigt. Dennoch lässt sich innerhalb des Musters ein stark konservierter Bereich aus 13 Aminosäuren von F135 bis P149 identifizieren, die direkt nebeneinander liegen. Auch in diesem Muster beschränken sich die Musterpositionen nicht auf den Bereich des aktiven Zentrums, sondern auch Positionen, die sich im Bereich des allosterischen Zenrums befinden, sind im Muster vorhanden. Einige laut Literatur wichtige Aminosäuren, z.B. D72, Y85, Y121, D141, E148 und K208, wurden im Muster eingetragen. Andere wichtige Aminosäuren, z.B. H143, S151, R 158, K160, R172 und Y254 sind im Muster nicht vertreten.

3.4.1.11 Vergleiche der erhaltenen Muster und Sequenzdomänen für EC 3.5.99.6 mit anderen Datenbanken

Wie den Tabellen 3-17 und 3-18 zu entnehmen ist, erstreckt sich die Proteindomäne für die Sequenzen des Clusterbaums 19905, E-Wert 10^{-59}, EC 3.5.99.6, auf die Sequenzabschnitte von etwa AS 66 bis AS 230. Die Proteindomäne des Clusterbaums 19905, E-Wert 10^{-94}, EC 3.5.99.6, reicht von etwa AS 1 bis AS 240.

Einträge für EC 3.5.99.6 in anderen Datenbanken:

CATH:

Einordnung der Domäne in die homologe Superfamilie (3.40.50.1360).
[PDB] Hydrolase.

Pfam:

Einordnung der Domäne in die Familie Glucosamine_iso (PF01182).

Glucosamine-6-phosphate isomerases/6-phosphogluconolactonase.
Die Domänenregion liegt bei der größten Sequenzgruppe zwischen AS 8 bis AS 230.

SCOP:

Fold: NagB/RpiA/CoA transferase-like.

Class: Alpha and beta proteins (a/b).

3.4.2 Beispiel 2:

L-Ribulose-5-Phosphat 4-Epimerase, EC 5.1.3.4, und L-Fuculose-Phosphat Aldolase, EC 4.1.2.17

In diesem Beispiel werden die Sequenzmuster, katalysierte Reaktionen und Eigenschaften der Enzyme L-Ribulose-5-Phosphat 4-Epimerase, EC 5.1.3.4 und L-Fuculose-Phosphat Aldolase, EC 4.1.2.17 behandelt. Dazu wird der Clusterbaum 18685 analysiert, in dem die Sequenzen der EC-Nummern EC 5.1.3.4 und EC 4.1.2.17 geclustert wurden. Zusätzlich befinden sich in diesem Clusterbaum auch die Sequenzen des Enzyms Rhamnulose-1-Phosphat Aldolase, EC 4.1.2.19, sowie Sequenzen der Enzyme Enoyl-CoA Hydratase, EC 4.2.1.17 und 3,4-Dihydroxyphthalat Decarboxylase, EC 4.1.1.69.

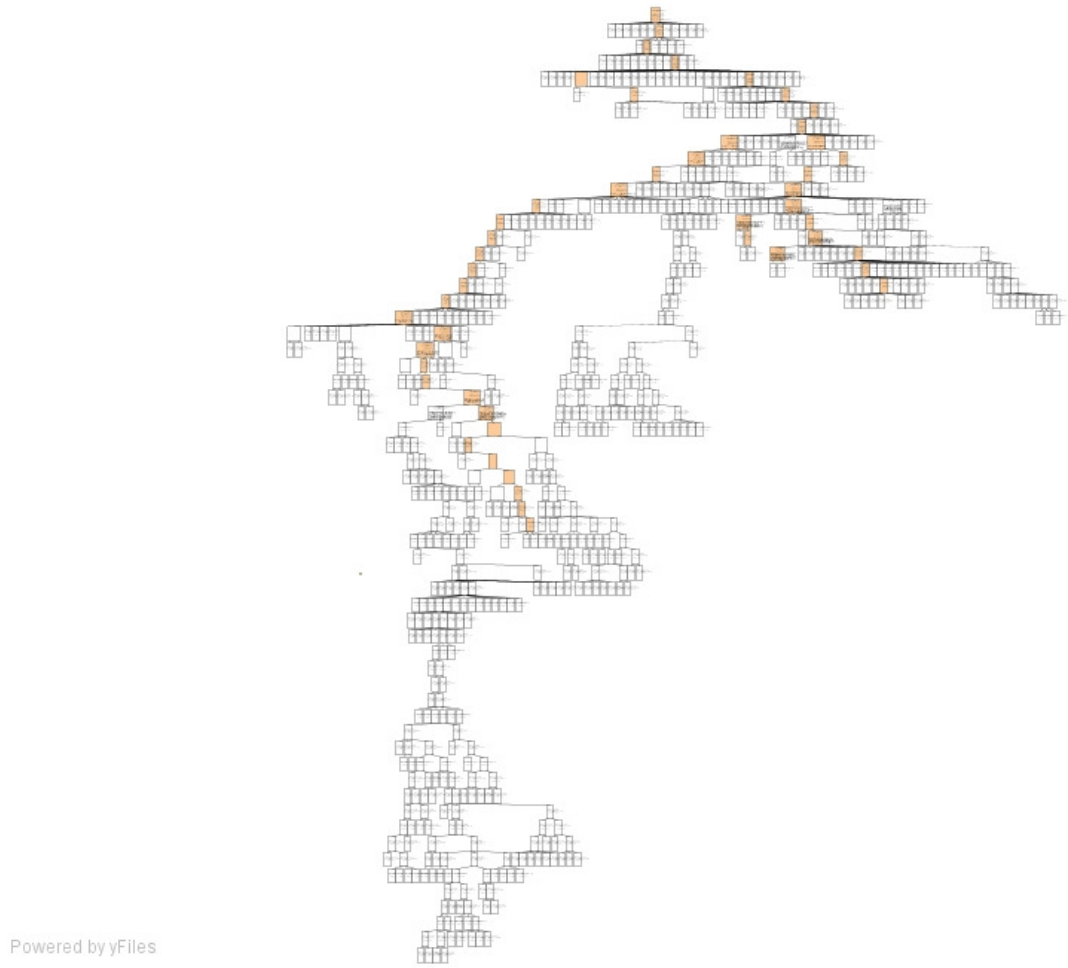


Abbildung 3-9: Der Clusterbaum 18685.

Von besonderem Interesse ist dieser Clusterbaum, da sich in der gesamten Datenbank kein anderer Clusterbaum befindet, indem Muster für die EC-Nummern EC 5.1.3.4, EC 4.1.2.17 und EC 4.1.2.19 erstellt wurden.

Die Tatsache, dass sich zwei EC-Klassen im Cluster 18685 befinden, führte zu der Entscheidung, in diesem Beispiel besonders auf den Vergleich der EC-Nummern EC 4.1.2.17 und EC 5.1.3.4 einzugehen.

Die folgende Tabelle 3-19 zeigt Details zu den Einträgen der Datenbank.

Cluster-baum	Node	E-Wert	Datenbank	Ec-Nummer	Sequenz-anzahl	Muster-länge	Richtig-Positive	Falsch-Positive
18685	36	10^{-6}	TRROT	4.1.2.17 4.1.2.19	6	74	6	0
18685	64	10^{-7}	TREMBL	4.1.2.17	2	201	2	3
18685	104	10^{-10}	TRROT	4.1.2.19	36	36	36	7
18685	142	10^{-13}	TRROT	4.1.2.17 4.2.1.17 5.1.3.4	97	16	108	150
18685	156	10^{-14}	TREMBL	4.1.2.17	2	115	2	0
18685	162	10^{-14}	TRROT	4.1.2.17 4.2.1.17 5.1.3.4	58	28	75	79
18685	167	10^{-14}	TRROT	4.1.2.17	26	30	26	20
18685	178	10^{-15}	TRROT	4.1.2.17 5.1.3.4	3	66	3	0
18685	202	10^{-16}	TRROT	4.1.2.17 4.2.1.17 5.1.3.4	47	36	61	59
18685	210	10^{-17}	TREMBL	4.1.2.17 5.1.3.4	2	69	2	0
18685	325	10^{-22}	TREMBL	4.1.2.17	2	166	2	0
18685	332	10^{-22}	TREMBL	4.1.2.17	8	59	8	22
18685	343	10^{-22}	TRROT	4.1.2.17 4.2.1.17 5.1.3.4	116	12	116	193
18685	361	10^{-24}	TRROT	4.1.2.17 4.2.1.17 5.1.3.4	114	14	115	164
18685	362	10^{-24}	TREMBL	5.1.3.4	4	90	4	2
18685	366	10^{-27}	TREMBL	4.1.2.17	2	116	2	0
18685	372	10^{-31}	TRROT	4.1.2.17 4.2.1.17 5.1.3.4	111	17	113	146
18685	375	10^{-33}	TRROT	5.1.3.4	69	38	69	67
18685	376	10^{-33}	TRROT	4.1.2.17 4.2.1.17	42	36	42	36
18685	395	10^{-37}	TREMBL	4.1.2.17	2	130	2	0
18685	397	10^{-37}	TRROT	4.1.2.17 4.2.1.17	40	43	40	29
18685	410	10^{-42}	TRROT	4.1.2.17	27	56	27	15

18685	433	10^{-50}	TREMBL	4.1.2.17 4.2.1.17	12	128	12	2
18685	493	10^{-61}	TREMBL	4.1.2.17	2	212	2	0
18685	496	10^{-61}	TREMBL	4.1.2.17 4.2.1.17	10	145	10	2
18685	519	10^{-68}	TREMBL	4.1.2.17	3	205	3	0

Tabelle 3-19: Übersicht über die generierten Muster des Clusterbaums 18685.

Der Clusterbaum 18685 besteht aus 735 Knoten und gehört damit zu den größeren Clusterbäumen der Datenbank. Es wurden 542 Sequenzen geclustert, davon besitzen 254 Sequenzen die EC-Nummer 5.1.3.4, 185 Sequenzen die EC-Nummer 4.1.2.17, 89 Sequenzen die EC-Nummer 4.1.2.19, 12 Sequenzen die EC-Nummer 4.2.1.17 und zwei Sequenzen die EC-Nummer 4.1.1.69. Es existieren im Clusterbaum 18685 zehn Muster für EC-Nummer 4.1.2.17, 2 Muster für EC 5.1.3.4, ein Muster für EC 4.1.2.19, vier Muster für EC 4.1.2.17 + EC 4.2.1.17, ein Muster für EC 4.1.2.17 + EC 4.1.2.19, sechs Muster für EC 4.1.2.17 + EC 4.2.1.17 + EC 5.1.3.4 und zwei Muster für EC 4.1.2.17 + EC 5.1.3.4.

In den folgenden Abschnitten werden besonders die Enzyme der EC-Nummern 5.1.3.4 und EC 4.1.2.17 strukturell und funktionell untersucht; mit besonderem Augenmerk auf die strukturelle Einordnung der Musterpositionen der erstellten Muster, sowie auf die Funktionen des Enzyms. Die Angaben für die Positionen der Aminosäuren beziehen sich bei beiden Enzymen auf die Sequenzen von *Escherichia coli.*

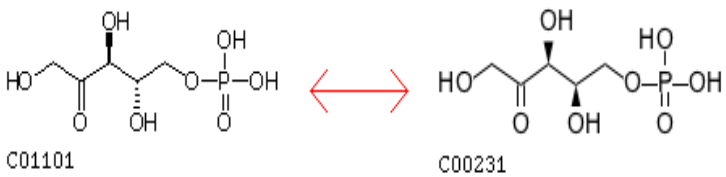
3.4.2.1 Eigenschaften von EC 5.1.3.4 und EC 4.1.2.17

EC-Nummer 5.1.3.4:

Name des Enzyms: L-Ribulose-5-Phosphat 4-Epimerase

Das Enzym L-Ribulose-5-Phosphat 4-Epimerase katalysiert die Epimerisierung von L-Ribulose-5-Phosphat zu D-Xylulose 5-Phosphat.

Reaktion: L-Ribulose 5-Phosphat ↔ D-Xylulose 5-Phosphat

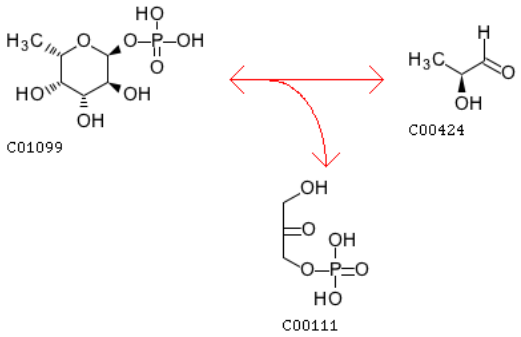


EC-Nummer 4.1.2.17:

Name des Enzyms: L-Fuculose-Phosphat Aldolase

Das Enzym L-Fuculose-Phosphat Aldolase gehört zu den Aldehyd-Lyasen und katalysiert die Umkehrung einer Aldol Kondensation.

Reaktion: L-Fuculose 1-Phosphat ↔ Glyceron Phosphat + (S)-Lactaldehyd



L-Fuculose-Phosphat Aldolase (FucA) gehört zur Klasse 2 Aldolasen, d.h. das Enzym benötigt während der Reaktion ein Metall-Ion, meist Zink, das Reaktionsintermediate stabilisiert. L-Fuculose-Phosphat Aldolase ist sehr spezifisch für Dihydroxyaceton Phosphat, kann aber verschiedene Aldehyde als Aldol-Akzeptor nutzen [153].

Die angegebene Reaktion bildet eine entscheidende Reaktion während des bakteriellen L-Fucose Metabolismus [154]. Im Gegensatz zu Klasse 1 Aldolasen, die während der Katalyse eine Schiffs-Base bilden und aus einem $(\alpha/\beta)_8$-Barrel bestehen, bildet L-Fuculose-Phosphat Aldolase in *E. coli* ein Homotetramer in C4 Symmetrie (206 Aminosäuren pro Untereinheit) [153]. Das Enzym hat ein Molekulargewicht von 23,8 kDa. In jeder Untereinheit befindet sich ein aktives Zentrum. Das Enzym fand in letzter Zeit großes Interesse, da es Zucker synthetisieren kann, die im Labor nur schwierig herzustellen sind [154]. L-Ribulose-5-Phosphat 4-Epimerase (L-Ru4P) ist ein Enzym, das Bakterien ermöglicht, Arabinose als Energiequelle zu nutzen. Es bildet ein Homotetramer aus vier identischen Untereinheiten, mit einem Molekulargewicht von jeweils 25,5 kDa [155]. Es wurde zunächst angenommen, dass der Reaktionsmechanismus dem von UDP-Galactose-4-Epimerase ähnelt [156]. Es zeigte sich aber, dass keine Protonierung/Deprotonierung stattfindet und dass kein NAD^+-Cofaktor nötig ist. Über Sequenzalignments wurde die Sequenzähnlichkeit zu Aldolasen der Klasse 2 erkannt.

Die Sequenzen der beiden besprochenen Enzyme sind zu großen Teilen identisch und sind höchstwahrscheinlich evolutionär miteinander verwandt [155]. Zudem verwenden beide Enzyme ein Metall-Ion während der Katalyse und bilden die gleiche Quartärstruktur. Auch ist das phosphatbindende aktive Zentrum in beiden Enzymen größtenteils identisch [157]. Veröffentlichungen sprechen in Zusammenhang von L-Ribulose-5-Phosphat 4-Epimerase von einer verdeckten Aldolase [156] und direkte Vergleiche zwischen beiden Enzymen wurden konstruiert [157].

3.4.2.2 Der katalytische Mechanismus von EC 4.1.2.17 und EC 5.1.3.4

Aufgrund der hohen Sequenzähnlichkeit zwischen den Enzymen L-Fuculose-Phosphat Aldolase und L-Ribulose-5-Phosphat 4-Epimerase und den damit verbundenen ähnlichen Reaktionsmechanismen werden die funktionellen Eigenschaften für beide Enzyme in diesem Abschnitt zusammengefasst.

Die Enzyme binden während der Katalyse wie beschrieben ein Metall-Ion im reaktiven Zentrum. Die ionenbindenden Aminosäuren sind bei L-Fuculose-Phosphat Aldolase E73, H92, H94 und H155. Entsprechende funktionelle Positionen sind auch bei L-Ribulose-5-Phosphat 4-Epimerase vorhanden: D76, H95, H97 und H171. Mutationsexperimente dieser Aminosäuren führten zu einer erheblichen Verschlechterung der Zink-Ionen Bindung und einer daraus resultierenden Reduzierung der Wechselzahl der Enzyme [156]. Die Aminosäuren T43, G44, S71 und S72 bei L-Fuculose-Phosphat Aldolase, entsprechend S44, G45, S74 und S75 bei L-Ribulose-5-Phosphat 4-Epimerase binden die Phosphatgruppe des Substrats [158]. In Abbildung 3-10 wurden die Aminosäuren der aktiven Zentren farbig markiert.

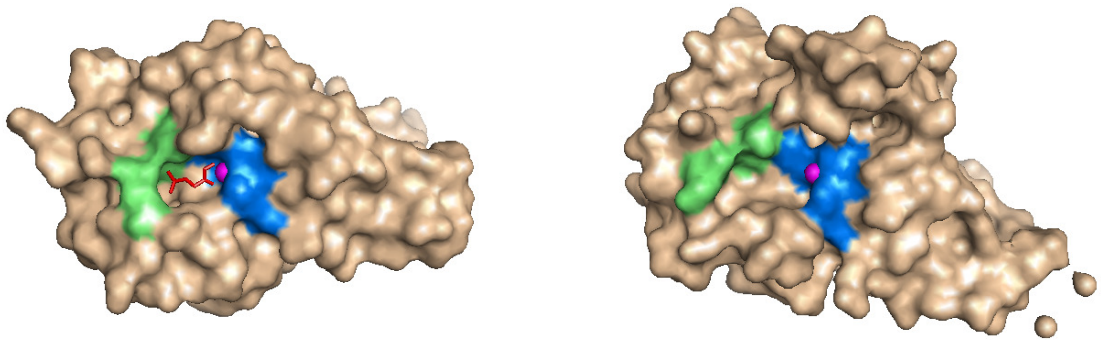


Abbildung 3-10: L-Fuculose-Phosphat Aldolase (PDB 4FUA) links, bindet Phosphoglycolohydroxamat (PGH), L-Ribulose-5-Phosphat 4-Epimerase (PDB 1K0W) rechts. Zink-Ion pink, Zink-bindende Aminosäuren marine, Phosphat-bindende Aminosäuren grün, jeweils Monomere, *E. coli.*

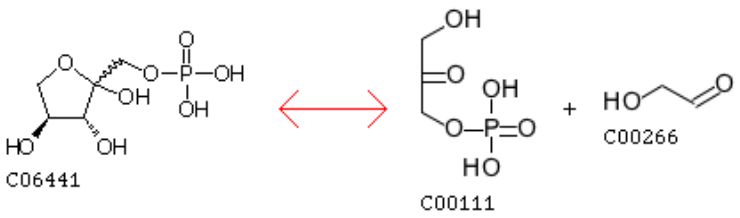
Die Mutationen N28A und K42M in L-Ribulose-5-Phosphat 4-Epimerase führen zu einem rapiden Anstieg des K_m-Werts, was die Bedeutung dieser Aminosäuren für die Bindung der Phosphatgruppe verdeutlicht [157]. H218 scheint eine wichtige strukturelle Aminosäure zu sein. H218 selbst bildet eine H-Brücke zu Y141. N28 ist in beiden Enzymen hochkonserviert (entspricht N29 in L-Fuculose-Phosphat Aldolase) und bildet eine H-Brücke mit dem Aldolase-Inhibitor Komplex. Die Mutante N28A hat wie erwartet eine verminderte Affinität zum Substrat [157]. K42 befindet sich unterhalb der Phosphatbindestelle und ist bei der Bindung des negativ geladenen Phosphats beteiligt. Eine dieser Aminosäure entsprechende Position gibt es in der Sequenz der L-Fuculose-Phosphat Aldolase nicht. Zu einer Stabilisierung des aktiven Zentrums trägt zudem E142 bei, die eine Salzbrücke zu R221 eingeht. Beide Positionen sind in der Aminosäuresequenz hochkonserviert [157].

3.4.2.3 Weitere EC-Nummern im Clusterbaum und deren katalysierte Reaktionen

EC-Nummer 4.1.2.19:

Name des Enzyms: L-Xylulose 1-Phosphat Lactaldehyd-Lyase

Reaktion: L-Xylulose 1-Phosphat ↔ Glycerone Phosphat + Glycolaldehyd

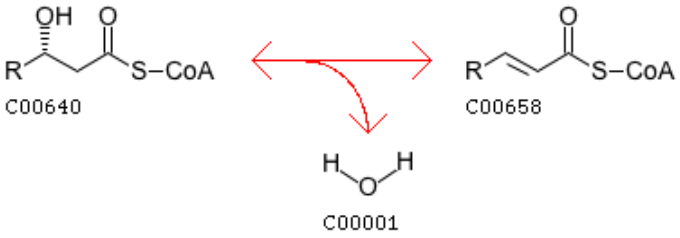


EC-Nummer 4.2.1.17:

Name des Enzyms: Enoyl-CoA Hydratase

Das Enzym katalysiert die Spaltung einer Kohlenstoff-Sauerstoff Bindung, bei der Wasser freigesetzt wird.

Reaktion: (3S)-3-Hydroxyacyl-CoA ↔ trans-2,3-Dehydroacyl-CoA + H_2O

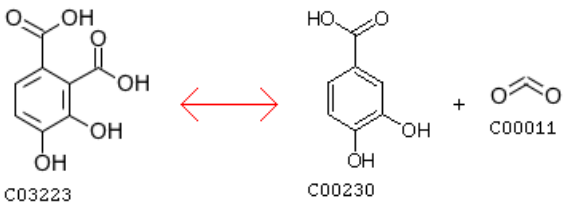


EC-Nummer 4.1.1.69:

Name des Enzyms: 3,4-Dihydroxyphthalat Decarboxylase

Das Enzym katalysiert die Decarboxylierung von 3,4-Dihydroxyphthalat. Dabei entstehen 3,4-Dihydrobenzoat und Kohlendioxid.

Reaktion: 3,4-Dihydroxyphthalat ↔ 3,4-Dihydroxybenzoat + CO_2



3.4.2.4 Größte gemeinsame Teilstrukturen der Edukte, Produkte und Co-Substrate der Enzymreaktionen

Die größten gemeinsamen Teilstrukturen der Edukte, Produkte und Co-Substrate werden für die Reaktionen der EC-Nummern EC 5.1.3.4 + EC 4.1.2.17, EC 5.1.3.4 + EC 4.1.1.69, EC 5.1.3.4 + EC 4.2.1.17, EC 4.1.2.17 + EC 4.2.1.17, EC 4.1.2.17 + EC 4.1.1.69 und EC 4.2.1.17 + EC 4.1.1.69 verglichen. Aufgrund der Vielzahl der möglichen EC-Kombinationen,

beschränkt sich die Darstellung der Ergebnisse in den Tabellen auf die Vergleiche, bei denen die zwei größten gemeinsamen Teilstrukturen ermittelt wurden.

Vergleich EC 5.1.3.4 EC 4.1.2.17	**Molfile**	**Atome**	**Molekül**
	12395.mol	14	L-Ribulose 5-Phosphat
	12185.mol	15	L-Fuculose 1-Phosphat
c-MCS		10	
	12185.mol	15	L-Fuculose 1-Phosphat
	9407.mol	14	D-Xylulose 5-Phosphat
c-MCS		14	

Tabelle 3-20: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 5.1.3.4 und EC 4.1.2.17.

Vergleich EC 5.1.3.4 EC 4.1.1.69	**Molfile**	**Atome**	**Molekül**
	12395.mol	14	L-Ribulose 5-Phosphat
	3281.mol	14	3,4-Dihydroxyphthalat
c-MCS		8	
	9407.mol	14	D-Xylulose 5-Phosphat
	3257.mol	11	3,4-Dihydroxybenzoat
c-MCS		7	

Tabelle 3-21: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 5.1.3.4 und EC 4.1.1.69.

Vergleich EC 5.1.3.4 EC 4.2.1.17	**Molfile**	**Atome**	**Molekül**
	12395.mol	14	L-Ribulose 5-Phosphat
	235.mol	54	(3S)-3-Hydroxyacyl-CoA
c-MCS		13	
	9407.mol	14	D-Xylulose 5-Phosphat
	235.mol	54	(3S)-3-Hydroxyacyl-CoA
c-MCS		13	

Tabelle 3-22: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 5.1.3.4 und EC 4.2.1.17.

Vergleich EC 4.1.2.17 EC 4.2.1.17	**Molfile**	**Atome**	**Molekül**
	12185.mol	15	L-Fuculose 1-Phosphat
	235.mol	54	(3S)-3-Hydroxyacyl-CoA
c-MCS		9	
	11085.mol	10	Glyceron Phosphat
	235.mol	54	(3S)-3-Hydroxyacyl-CoA
c-MCS		10	

Tabelle 3-23: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 4.1.2.17 und EC 4.2.1.17.

Vergleich EC 4.1.2.17 EC 4.1.1.69	**Molfile**	**Atome**	**Molekül**
	12185.mol	15	L-Fuculose 1-Phosphat
	3281.mol	14	3,4-Dihydroxyphthalat
c-MCS		9	
	463.mol	5	(S)-Lactaldehyd
	3281.mol	14	3,4-Dihydroxyphthalat
c-MCS		5	

Tabelle 3-24: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 4.1.2.17 und EC 4.1.1.69.

Vergleich EC 4.2.1.17 EC 4.1.1.69	Molfile	Atome	Molekül
	3281.mol	14	3,4-Dihydroxyphthalat
	235.mol	54	(3S)-3-Hydroxyacyl-CoA
c-MCS		8	
	8725.mol	3	CO_2
	235.mol	54	(3S)-3-Hydroxyacyl-CoA
c-MCS		2	

Tabelle 3-25: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 4.2.1.17 und EC 4.1.1.69.

Wie den Tabellen 3-20 bis 3-25 entnommen werden kann, sind die Edukte, Produkte oder Co-Substrate der untersuchten Enzymreaktionen sehr ähnlich.

3.4.2.5 Vergleich der EC-Kombinationen mit der Clusterung nach identischen R-Strings

Die Enzyme mit den EC Nummern EC 4.1.2.- und EC 4.1.1- wurden aufgrund gleicher R-Strings gruppiert (vgl. Abschnitte 2.10 und 2.10.4, Tabelle 2-3). Stellvertretend für diese EC-Subsubklassen wird die Reaktionsmatrix von Ketotetrose-Phosphat Aldolase, EC 4.1.2.2 und Pyruvat Decarboxylase, EC 4.1.1.1, dargestellt. Die Reaktionsmatrix ist für beide Reaktionen identisch. Eine identische Reaktionsmatrix bedeutet, dass das Elektronentransfermuster identisch ist und gleiche Bindungen auf der Edukt- und Produktseite gespalten werden. Der Reaktionskern ist identisch.

Reaktionsmatrix für EC 4.1.2.- und EC 4.1.1.-:

	O	H	C	C
O	0	−1	0	1
H	−1	0	1	0
C	0	1	0	−1
C	1	0	−1	0

Beide Enzyme katalysieren die Spaltung von jeweils einer C-C und O-H Bindung, sowie den Aufbau der Atombindungen C-H und C-O.

3.4.2.6 Untersuchung von Falsch-Positiven Treffern ausgewählter Sequenzmuster

In diesem Abschnitt wird das Muster für EC 5.1.3.4, Clusterbaum 18685, E-Wert 10^{-33}, untersucht. Es wurde aus 69 Sequenzen generiert und hat bei der Suche gegen SWISS-PROT

und TrEMBL 69 Richtig-Positive und 67 Falsch-Positive Treffer erzielt. Die folgende Tabelle zeigt Einzelheiten zu allen Falsch-Positiven Treffer des Musters.

Sequenz	Start	Ende	Datenbank	Bezeichnung
Q08PR0_STIAU	59	203	TREMBL	Putative epimerase-aldolase
Q1U7L3_LACRE	56	202	TREMBL	Class II aldolase/adducin-like
Q1XUQ6_CYTJO	59	205	TREMBL	Class II aldolase/adducin-like
Q2BS95_LACRE	56	202	TREMBL	L-ribulose 5-phosphate 4-epimerase
Q3DDB5_STRAG	59	205	TREMBL	Carbohydrate isomerase,
Q0T9J5_ECOL5	55	199	TREMBL	Probable sugar isomerase SgaE
Q2WQD3_CLOBE	56	200	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q3D012_STRAG	59	205	TREMBL	Carbohydrate isomerase
A0KWX6_SHESA	57	201	TREMBL	Class II aldolase/adducin family
A0NIT5_OENOE	57	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
A1ES06_VIBCH	56	202	TREMBL	Sugar isomerase SgaE
A1F4V4_VIBCH	56	202	TREMBL	Sugar isomerase SgaE
A1JM10_YERE8	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q0GLA8_LACRE	56	202	TREMBL	L-ribulose 5-phosphate 4-epimerase
Q1R509_ECOUT	56	203	TREMBL	Probable sugar isomerase SgbE
Q3DL50_STRAG	59	205	TREMBL	Carbohydrate isomerase
A0UXX5_CLOCE	56	200	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q03HQ1_PEDPA	56	202	TREMBL	Ribulose-5-phosphate 4-epimerase
Q034C7_LACC3	56	202	TREMBL	Ribulose-5-phosphate 4-epimerase
Q9LBQ2_MYCSM	56	191	TREMBL	L-ribulose-5-phosphate-4-epimerase
A1XPK6_KLEPN	56	203	TREMBL	YiaS
Q03PR4_LACBA	56	202	TREMBL	Ribulose-5-phosphate 4-epimerase
Q0T8D6_SHIF8	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q3D7L1_STRAG	59	205	TREMBL	Carbohydrate isomerase
A1SDN9_NOCSJ	67	202	TREMBL	Class II aldolase/adducin family protein
Q0SX87_SHIF8	55	199	TREMBL	Putative epimerase/aldolase
Q0TTD9_CLOP1	56	200	TREMBL	Putative L-ribulose 5-phosphate 4-epimerase SgaE
Q8D0J2_YERPE	92	239	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q9X6B7_YERPE	81	228	TREMBL	L-ribulose-phosphate 4-epimerase
A2PGP2_VIBCH	56	202	TREMBL	Sugar isomerase SgaE, AraD/FucA family
A2P940_VIBCH	56	202	TREMBL	Sugar isomerase SgaE, AraD/FucA family
A2UGU0_ECOLI	55	199	TREMBL	Class II aldolase/adducin family protein
A3EG34_VIBCH	56	202	TREMBL	Sugar isomerase SgaE, AraD/FucA family
A3EKW3_VIBCH	56	202	TREMBL	Sugar isomerase SgaE, AraD/FucA family
Q326H4_SHIBS	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q32K32_SHIDS	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q3Z5V0_SHISS	56	233	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q8DXP0_STRA5	59	205	TREMBL	Carbohydrate isomerase, AraD/FucA family
Q7MBY1_VIBVY	56	202	TREMBL	Ribulose-5-phosphate 4-epimerase
Q8P2S7_STRP8	80	226	TREMBL	Putative L-ribulose 5-phosphate
Q57TG0_SALCH	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q97JE5_CLOAB	56	202	TREMBL	Ribulose-5-phosphate 4-epimerase family protein
Q9KMS6_VIBCH	56	202	TREMBL	Sugar isomerase SgaE
Q5PLP5_SALPA	56	203	TREMBL	Putative sugar isomerase
Q83HE2_TROW8	52	185	TREMBL	Putative sugar isomerase
Q8D544_VIBVU	56	202	TREMBL	Ribulose-5-phosphate 4-epimerase
Q98PX0_MYCPU	64	211	TREMBL	SUGAR ISOMERASE SGAE
Q5PDF3_SALPA	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q5PXF3_SALPA	57	204	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q8E3B0_STRA3	59	205	TREMBL	Hypothetical protein gbs1851
Q9CLI6_PASMU	68	215	TREMBL	AraD
A1A7A8_ECOK1	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
A1AJA3_ECOK1	55	199	TREMBL	Probable sugar isomerase SgaE
Q0T9J5_ECOL5	55	199	TREMBL	Probable sugar isomerase SgaE
Q1R509_ECOUT	56	203	TREMBL	Probable sugar isomerase SgbE
Q4A7T5_MYCH7	63	210	TREMBL	Sugar isomerase SgaE

Q8FCC7_ECOL6	56	203	TREMBL	Sugar isomerase SgaE
Q1R509_ECOUT	56	203	TREMBL	Probable sugar isomerase sgbE
Q1R362_ECOUT	71	215	TREMBL	Probable sugar isomerase SgaE
Q0TBL7_ECOL5	56	203	TREMBL	Probable L-ribulose-5-phosphate 4-epimerase
Q97NJ4_STRPN	52	198	TREMBL	Putative L-ribulose 5-phosphate 4-epimerase AraD
Q0GLA8_LACRE	56	202	TREMBL	L-ribulose 5-phosphate 4-epimerase
Q32K32_SHIDS	56	203	TREMBL	L-ribulose-5-phosphate 4-epimerase
Q3D7L1_STRAG	59	205	TREMBL	Carbohydrate isomerase
Q034C7_LACC3	56	202	TREMBL	Ribulose-5-phosphate 4-epimerase
Q3D7L1_STRAG	59	205	TREMBL	Carbohydrate isomerase
P44989	56	203	SPROT	Probable sugar isomerase sgbE

Tabelle 3-26: Falsch-Positive Treffer für EC 5.1.3.4 aus Clusterbaum 18685, E-Wert 10^{-33}.

Die folgende Tabelle gibt das Ergebnis der Untersuchung des Musters für EC 4.1.2.17, Clusterbaum 18685, E-Wert 10^{-14} auf Falsch-Positive Treffer wieder. Das Muster wurde aus 26 Sequenzen generiert und hat bei der Suche gegen SWISS-PROT und TrEMBL 26 Richtig-Positive und 20 Falsch-Positive Treffer erzielt. Es existieren keine Falsch-Positiven Treffer bei der Suche nach Treffern des Musters gegen die Datenbank SWISS-PROT. Die Ergebnisse aller Falsch-Positiven Treffer sind in folgender Tabelle zusammengefasst.

Sequenz	**Start**	**Ende**	**Datenbank**	**Bezeichnung**
A1BAW5_PARDP	12	101	TREMBL	Transcriptional regulator, Fis family
Q2DDU5_ACICY	10	102	TREMBL	Putative L-fuculose phosphate aldolase protein
Q1YJP3_9RHIZ	12	104	TREMBL	Aldolase
Q05W61_9SYNE	10	101	TREMBL	Putative L-fuculose phosphate aldolase protein
Q0G1R2_9RHIZ	12	104	TREMBL	Putative l-fuculose phosphate aldolase protein
Q358C8_9BRAD	12	104	TREMBL	Putative L-fuculose phosphate aldolase protein
Q0FXG5_9RHIZ	12	104	TREMBL	Putative l-fuculose phosphate aldolase protein
Q0T159_SHIF8	10	99	TREMBL	L-fuculose-1-phosphate aldolase
A0NYY9_9RHOB	16	107	TREMBL	L-fuculose phosphate aldolase
Q2NWL7_SODGM	10	99	TREMBL	L-fuculose-1-phosphate aldolase
Q3J0N6_RHOS4	12	101	TREMBL	L-fuculose-phosphate aldolase
Q57KE3_SALCH	10	99	TREMBL	L-fuculose-1-phosphate aldolase
Q3YY56_SHISS	10	99	TREMBL	L-fuculose-1-phosphate aldolase
Q92W73_RHIME	15	103	TREMBL	Putative L-fuculose phosphate aldolase protein
Q7MJI5_VIBVY	10	99	TREMBL	Ribulose-5-phosphate 4-epimerase
Q31XJ1_SHIBS	10	99	TREMBL	L-fuculose-1-phosphate aldolase
A1AEY7_ECOK1	10	99	TREMBL	L-fuculose phosphate aldolase
Q32CB8_SHIDS	10	99	TREMBL	L-fuculose-1-phosphate aldolase
Q89H25_BRAJA	12	101	TREMBL	L-fuculose phosphate aldolase
Q8Z431_SALTI	10	99	TREMBL	Fuculose-1-phosphate aldolase

Tabelle 3-27: Falsch-Positive Treffer für EC 4.1.2.17 aus Clusterbaum 18685, E-Wert 10^{-14}.

Aus den Bezeichnungen der getroffenen Sequenzen (vgl. Tabellen 3-26 und 3-27) kann geschlossen werden, dass zumeist Sequenzen der Enzyme L-Ribulose-5-Phosphat 4-Epimerase und L-Fuculose-Phosphat Aldolase getroffen werden. Diese wurde als Falsch-Positiv bewertet, da keine EC-Nummern angegeben wurden. Eine Übersicht der Kriterien, die für Beurteilung von Treffern maßgebend sind, liefert der Abschnitt 2.9 im Teil Daten, Algorithmen und Methoden.

3.4.2.7 PROSITE- und Clustermuster

Für die Enzyme L-Ribulose-5-Phosphat 4-Epimerase, EC 5.1.3.4, sowie für L-Fuculose-Phosphat Aldolase, EC 4.1.2.17, sind in der PROSITE-Datenbank keine Einträge vorhanden (Stand April 2008).

Folgende Abbildung zeigt das Clustermuster für EC 5.1.3.4 + EC 4.1.2.17 + EC 4.2.1.17, Clusterbaum 18685, E-Wert 10^{-13}.

```
[ILMV]-x(4)-[IL]-x(5)-G-N-x(10,14)-[ILV]-x-P-[ST]-x(11)-[ILMV]-x(13,20)-P-S-
[ST]-[DE]-x(3)-[HY]-x(12)-[ILV]-x-H-x-H
```

Abbildung 3-11: Muster für EC 5.1.3.4 + EC 4.1.2.17 + EC 4.2.1.17, Clusterbaum 18685, E-Wert 10^{-13}.

Folgende Abbildung zeigt das Clustermuster für EC 5.1.3.4, Clusterbaum 18685, E-Wert 10^{-33}. Die Musterpositionen, die sich im Clustermuster bei E-Wert 10^{-13} erhalten haben, wurden farbig markiert.

```
[ILM]-[IV]-x(3)-[FILM]-x(4,5)-[ILV]-x(5,19)-P-S-S-D-x(2)-[AST]-[HY]-x(2)-
[ILV]-Y-x(8)-[IV]-x-H-T-H-[AS]-x(2)-[AS]-x(2)-[FWY]-[AS]-x(5)-[ILV]-x(4)-T-
x(3)-D-x(5)-[IV]-P-x(6,9)-[IV]-x(5,8)-G-x(2)-I-x(3)-[FIL]-x(6,13)-[ILV]-
[IMV]-x(2)-H-x(2)-F-x(8,9)-[AS]-[ILV]-x(2)-[AS]-x(3)-E-x(11)-[ILM]
```

Abbildung 3-12: Muster für EC 5.1.3.4, Clusterbaum 18685, E-Wert 10^{-33}.

Das Clustermuster für EC 4.1.2.17, Clusterbaum 18685, E-Wert 10^{-14}. Musterpositionen, die sich im Clustermuster bei E-Wert 10^{-13} erhalten haben, wurden farbig markiert.

```
[ILM]-x(6)-[LMV]-x(4)-[IL]-x(3)-[AT]-[AST]-G-N-[ILV]-x(2)-R-x(4,8)-[FILM]-x-
[ILV]-[ST]-P-[ST]-x(6)-[ILM]-x(4)-[ILMV]-x(2)-[ILMV]-x(10,20)-P-S-[ST]-E-W-
x-[FLM]-H-x(6)-[KR]-x-[DE]-x(3)-[IV]-x-H-x-H-x(4)-[AST]
```

Abbildung 3-13: Muster für EC 4.1.2.17, Clusterbaum 18685, E-Wert 10^{-14}.

3.4.2.8 Identifizierungen der generierten Muster in den 3D-Darstellungen der Enzyme

Das resultierende Muster des Clusters bei E-Wert 10^{-33}, das aus 69 Sequenzen generiert wurde (EC 5.1.3.4) und das Muster des Clusters bei E-Wert 10^{-14}, das aus 26 Sequenzen generiert wurde (EC 4.1.2.17) bestehen aus 38 bzw. 30 Positionen und sind damit im Vergleich zu anderen Mustern der Datenbank relativ kurz. Das Muster für EC 5.1.3.4 + EC 4.1.2.17 + EC 4.2.1.17, E-Wert 10^{-13}, wurde aus 97 Sequenzen generiert und ist mit 16 Musterpositionen etwa um die Hälfte kürzer im Verglich zu den Mustern für EC 5.1.3.4 und EC 4.1.2.17.

Die Sequenz für das Enzym L-Fuculose-Phosphat Aldolase, EC 4.1.2.17, besteht in *E. coli* aus 206 Aminosäuren. Das generierte Muster für EC 4.1.2.17 aus dem Clusterbaum 18685 liegt im ersten Abschnitt der Sequenz im Bereich I10 bis T99. Die Musterpositionen wurden in der 3D-Struktur PDB Datei 1FUA farbig markiert und in Abbildung 3-14 (links) dargestellt. Aus dieser Abbildung geht hervor, dass sich die Musterpositionen stark auf das aktive Zentrum des Enzyms konzentrieren. Die Aminosäuren, die das aktive Zentrum bilden, sind stark konserviert. Während der Katalyse der chemischen Reaktion wird ein Metall-Ion gebunden. Die Aminosäuren E73, H92, H94 und H155 sind bei der Bindung des Metall-Ions beteiligt (vgl. Abschnitt 3.4.2.2). Im generierten Muster sind die Aminosäuren E73, H92 und H94 vorhanden. Die im Muster vorhandenen Aminosäuren T43, S71 und S72 binden während der Katalyse die Phosphatgruppe des Substrats (vgl. Abschnitt 3.4.2.2).

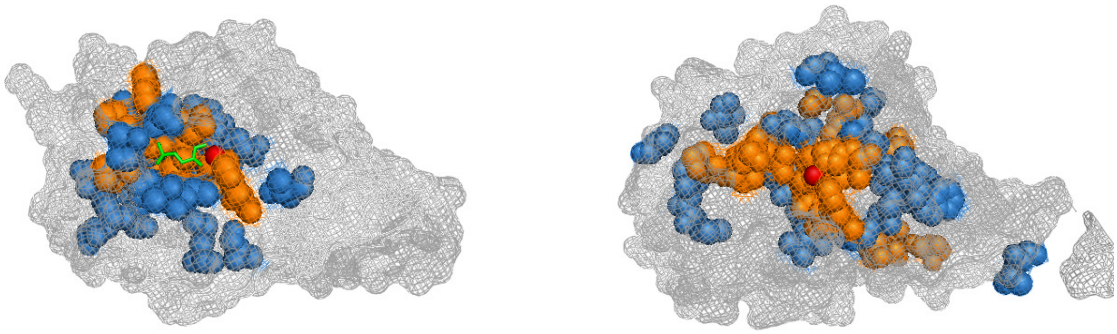


Abbildung 3-14: Links: L-Fuculose-Phosphat Aldolase, (PDB 4FUA) mit markierten Positionen des Musters aus Clusterbaum 18685, E-Wert 10^{-14} . Rechts: L-Ribulose-5-Phosphat 4-Epimerase, (PDB 1K0W) Muster aus Clusterbaum 18685, E-Wert 10^{-33}. Hochkonservierte Positionen orange, konservierte Positionen blau, Zink-Ion rot, zusätzlich in 4FUA: PGH grün. Jeweils Monomere, *E. coli.*

Auf der rechten Seite der Abbildung 3-14 wird die 3D-Struktur des Enzyms L-Ribulose-5-Phosphat 4-Epimerase, EC 5.1.3.4, PDB Datei 1K0W, aus *E. coli,* dargestellt. Die Sequenz besteht aus 223 Aminosäuren, das generierte Muster für EC 5.1.3.4 liegt im Sequenzbereich M56 bis L201. Wie schon das Beispiel für EC 4.1.2.17 zeigt, konzentrieren sich auch hier die Musterpositionen auf den Bereich des aktiven Zentrums. Das Muster ist im Vergleich zu dem Muster für EC 4.1.2.17 etwas länger. Vereinzelte Aminosäuren befinden sich auch außerhalb des aktiven Zentrums. Wie gut zu erkennen ist, sind Aminosäuren, die in direktem Kontakt zum Metall-Ion stehen, hochkonserviert. Hervorzuheben sind hierbei die Aminosäuren D76, H95, H97 und H171, die laut Fachliteratur bei der Bindung des Metall-Ions beteiligt sind (vgl. Abschnitt 3.4.2.2). Aminosäuren, die sich außerhalb dieses Bereichs befinden, sind funktionell konserviert. Die Aminosäuren S74 und S75 binden während der

Katalyse die Phosphatgruppe des Substrats. Beide Positionen sind hochkonserviert im Muster vorhanden.

Abbildung 3-15 zeigt die Sequenzmuster des Clusterbaums 18685 für die Enzyme Fuculose-Phosphat Aldolase, EC 4.1.2.17 (links) und L-Ribulose-5-Phosphat 4-Epimerase (rechts), die bei E-Wert 10^{-13} generiert wurden (vgl. Tabelle 3-19). Die Sequenzmuster wurden in den jeweiligen 3D-Strukturen farbig markiert.

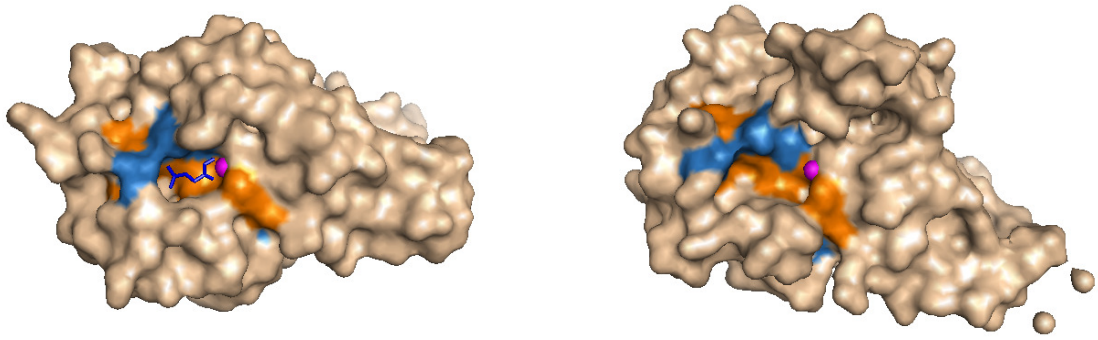


Abbildung 3-15: L-Fuculose-Phosphat Aldolase (PDB 4FUA) links, L-Ribulose-5-Phosphat 4-Epimerase (PDB 1K0W) rechts. Muster des Clusters bei E-Wert 10^{-13}. Hochkonservierte Positionen orange, konservierte Positionen marine, Zink-Ion pink, jeweils Monomere, *E. coli.*

Das Sequenzmuster für EC 4.1.2.17 und EC 5.1.3.4 hat eine Länge von 16 Aminosäuren und liegt im Bereich von M17 bis H94 auf der Sequenz des Enzyms L-Fuculose-Phosphat Aldolase und von L16 bis H97 auf der Sequenz des Enzyms L-Ribulose-5-Phosphat 4-Epimerase. In den Abbildungen ist die Oberfläche der Moleküle zu sehen, um zu verdeutlichen, dass vor allem Aminosäuren, die an der Bildung des aktiven Zentrums beteiligt sind, konserviert sind. Die „Tasche“ des aktiven Zentrums ist gut zu erkennen. Vor allem die Zink-Ion bindende Aminosäuren sind hochkonserviert. Im Vergleich zu den Mustern, die für die genannten EC-Nummern bei den E-Werten 10^{-14} und 10^{-33} generiert wurden, ist dieses Muster um 14 bzw. 22 Positionen kürzer. Als Konsequenz dieser Kürzung konzentrieren sich die Aminosäuren des Musters stärker auf das aktive Zentrum, als die zuvor generierten Muster, die bei kleineren E-Werten generiert wurden. Die Aminosäuren E73, H92 und H94 in L-Fuculose-Phosphat Aldolase, entprechend D76, H95 und H97 in L-Ribulose-5-Phosphat 4-Epimerase, die das Zink-Ion während der Katalyse binden, sind im Sequenzmuster vorhanden. Ebenso sind die Aminosäuren T43, S71, und S72 in L-Fuculose-Phosphat Aldolase, entsprechend S44, S74 und S75 in L-Ribulose-5-Phosphat 4-Epimerase, die während der chemischen Reaktion die

Phosphatgruppe des Substrats binden, als konservierte Positionen im generierten Sequenzmuster vorhanden.

3.4.2.9 Vergleiche der erhaltenen Muster und Sequenzdomänen für EC 5.1.3.4 und EC 4.1.2.17 mit anderen Datenbanken

Wie den Tabellen 3-26 und 3-27 zu entnehmen ist, erstreckt sich die Proteindomäne für die Sequenzen des Clusterbaums 18685, E-Wert 10^{-33}, EC 5.1.3.4, auf die Sequenzabschnitte von Aminosäure 56 bis Aminosäure 202. Die Proteindomäne des Clusterbaums 18685, E-Wert 10^{-14}, EC 4.1.2.17, reicht von circa Aminosäure 10 bis Aminosäure 100.

Einträge für EC 5.1.3.4 in anderen Datenbanken:

CATH:

Einordnung der Domäne in die homologe Superfamilie L-Fuculose-1-Phosphat Aldolase.

Pfam:

Einordnung der Domäne in Familie Aldolase_II and Adducin N-terminal domain (PF00596).

Bei 1051 Sequenzen liegt die Aldolase II Domäne zwischen den Aminosäuren 125 bis AS 312, bei 43 Sequenzen zwischen den Aminosäuren 20 bis AS 226.

SCOP:

Fold: AraD-like aldolase/epimerase.

Class: Alpha and beta proteins (a/b).

PANTHER:

Einordnung in die Familie FUCULOSE PHOSPHATE ALDOLASE (PTHR22789).

Einträge für EC 4.1.2.17 in anderen Datenbanken:

Die Einträge in CATH und Pfam entsprechen den oben dargestellten Ergebnissen der Suche für EC 5.1.3.4. PANTHER ordnet die Domänen in die Familie PENTULOSE-5-PHOSPHATE-4-EPIMERASE-RELATED (PTHR10640) ein.

3.4.3 Beispiel 3:

Carbamoylphosphat Synthase (glutamine-hydrolysing), EC 6.3.5.5, Aspartat Carbamoyltransferase, EC 2.1.3.2 und Carbamoylphosphat Synthase (ammoia), EC 6.3.4.16

In diesem Beispiel werden die Enzyme Carbamoylphosphat Synthase (glutamine-hydrolysing), EC 6.3.5.5, Aspartat Carbamoyltransferase, EC 2.1.3.2 und Carbamoylphosphat Synthase (ammoia), EC 6.3.4.16, besprochen. Dazu wird der Clusterbaum 31176 analysiert, in dem die EC-Nummern EC 6.3.5.5, EC 2.1.3.2 und EC 6.3.4.16 geclustert wurden. Zusätzlich befinden sich in diesem Clusterbaum auch die Sequenzen des Enzyms Phosphoribosylformylglycinamidine Synthase, EC 6.3.5.3. Auf die Bedeutung dieser Sequenzen wird in diesem Beispiel nicht näher eingegangen, da sich die EC-Nummer der Phosphoribosylformylglycinamidine Synthase nur in der laufenden Nummer im Vergleich zum Enzym Carbamoylphosphat Synthase unterscheidet. Von besonderem Interesse ist dieser Clusterbaum, da im Clusterbaum unterschiedliche EC-Klassen geclustert werden.

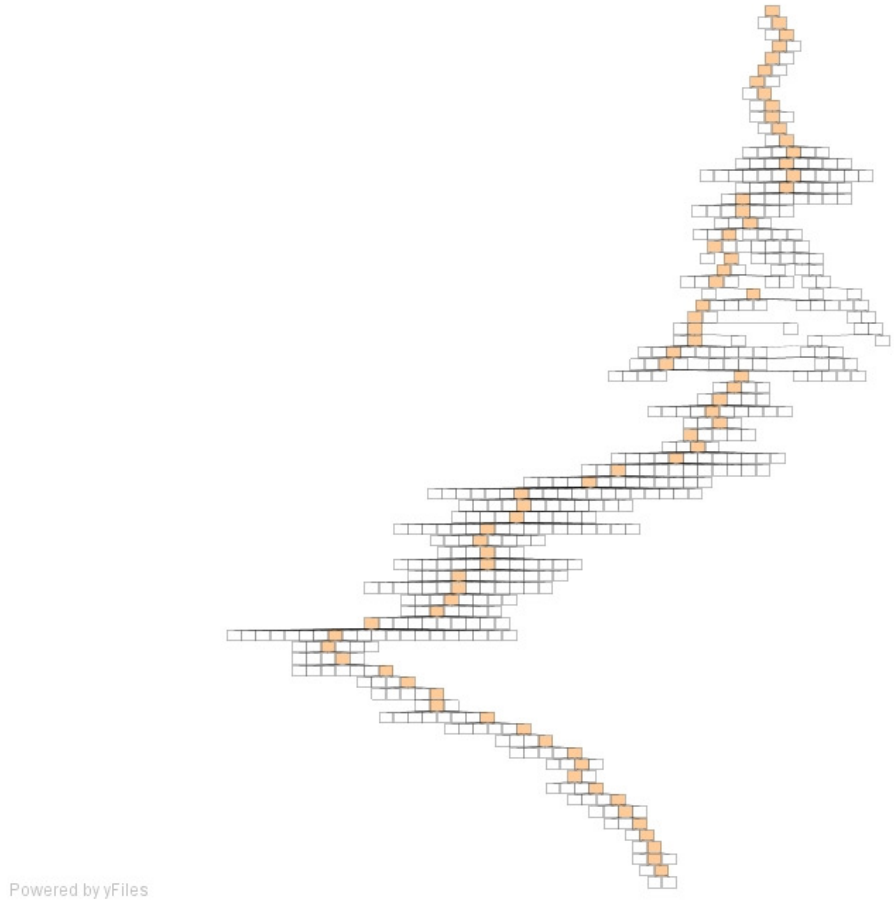


Abbildung 3-16: Der Clusterbaum 31176.

Der Clusterbaum 31176 besteht aus 482 Knoten und gehört damit zu den größeren Clusterbäumen der Datenbank. Bei E-Wert 10^{-2} wurden 382 Sequenzen geclustert, davon besitzen 368 Sequenzen die EC-Nummer 6.3.5.5 und 12 Sequenzen die EC-Nummer 6.3.4.16. Die EC-Nummern 2.1.3.2 und EC 6.3.5.3 sind mit jeweils einer Sequenz vertreten. Es existiert im Clusterbaum 31176, bei E-Wert 10^{-84}, ein Muster für die EC-Nummer 6.3.5.5, das aus 24 Sequenzen generiert wurde. Dieses Muster erzielte bei der Suche gegen SWISS-PROT und TrEMBL 43 Richtig-Positive und 19 Falsch-Positive Treffer.

Die folgende Tabelle zeigt Details zu den Einträgen in der Datenbank für Clusterbaum 31176. Sequenzen mit mehr als eine EC-Nummer pro Sequenz werden nicht dargestellt.

Cluster-Baum	Node	E-Wert	Datenbank	EC-Nummer	Sequenz-anzahl	Muster-länge	Richtig-Positive	Falsch-Positive
31176	1	10^{-2}	TRROT	6.3.5.3 2.1.3.2 6.3.5.5 6.3.4.16	311	75	312	346
31176	77	10^{-83}	TRROT	6.3.5.3 2.1.3.2 6.3.5.5 6.3.4.16	310	75	312	346
31176	84	10^{-84}	TRROT	6.3.5.5	24	210	43	19
31176	170	10^{-135}	TRROT	6.3.5.3 2.1.3.2 6.3.5.5 6.3.4.16	267	118	278	306
31176	175	10^{-136}	TRROT	2.1.3.2 6.3.5.5 6.3.4.16	264	119	277	309
31176	187	10^{-138}	TRROT	2.1.3.2 6.3.5.5 6.3.4.16	251	126	271	300
31176	195	10^{-139}	TRROT	2.1.3.2 6.3.5.5 6.3.4.16	248	126	271	304
31176	279	10^{-147}	TRROT	2.1.3.2 6.3.5.5 6.3.4.16	168	137	221	245
31176	291	10^{-148}	TRROT	2.1.3.2 6.3.4.16	152	138	216	241
31176	305	10^{-149}	TRROT	6.3.5.5 6.3.4.16	145	140	212	234
31176	435	10^{-164}	TRROT	6.3.5.5 6.3.4.16	34	212	38	47
31176	441	10^{-165}	TRROT	6.3.5.5 6.3.4.16	29	214	34	29
31176	473	10^{-176}	TRROT	6.3.5.5 6.3.4.16	6	372	6	0
31176	475	10^{-177}	TREMBL	6.3.5.5 6.3.4.16	5	404	5	0
31176	477	10^{-178}	TREMBL	6.3.5.5	3	437	3	0

Tabelle 3-28: Übersicht über die generierten Muster des Clusterbaums 31176.

3.4.3.1 Eigenschaften von EC 6.3.5.5

EC-Nummer 6.3.5.5:

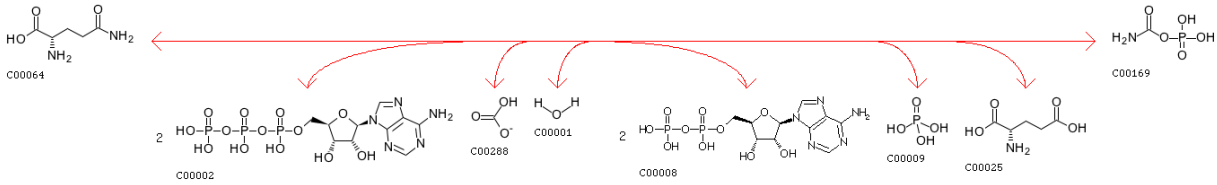
Name des Enzyms: Carbamoylphosphat Synthase (glutamine-hydrolysing)

Das Enzym Carbamoylphosphat Synthase gehört zur EC-Klasse der Ligasen und katalysiert die Bildung von Kohlenstoff-Stickstoff Bindungen.

Reaktion: 2 ATP + L-Glutamin + HCO_3- + H_2O

↔

2 ADP + Orthophosphat +L-Glutamat + Carbamoylphosphat



Carbamoylphosphat Synthase (CPS) kommt in den meisten lebenden Organsimen vor und katalysiert Carbamoylphosphat, die wichtig bei der Pyrimidin- und Argininsynthese ist, sowie im Harnstoffzyklus von Landtieren eine bedeutende Rolle spielt [159]. In *E. coli* bildet das Enzym ein α/β-Heterodimer aus einer kleinen Glutaminase- (~42 kDa) und einer großen ~118 kDa Synthetase Untereinheit [160]. Die kleine monofunktionale α-Untereinheit gehört zum Typ 1 der Amidotransferasen und katalysiert die Hydrolyse von Glutamin zu Glutamat und Ammoniak. Die große bifunktionale Untereinheit besitzt zwei ATP-Bindestellen [160]. Sie katalysiert im N-terminalen Teil die Bildung von Carbamat und im C-terminalen Teil die Phosphorylierung von Carbamat und Bicarbonat [161]. Das Enzym Carbamoylphosphat Synthase besitzt allosterische Zentren. Besonders hervorzuheben ist die Existenz von zwei intramolekularen Tunneln, die die entfernten aktiven Zentren untereinander verbinden, um Reaktionsintermediate vom Lösungsmittel zu schützen [161].

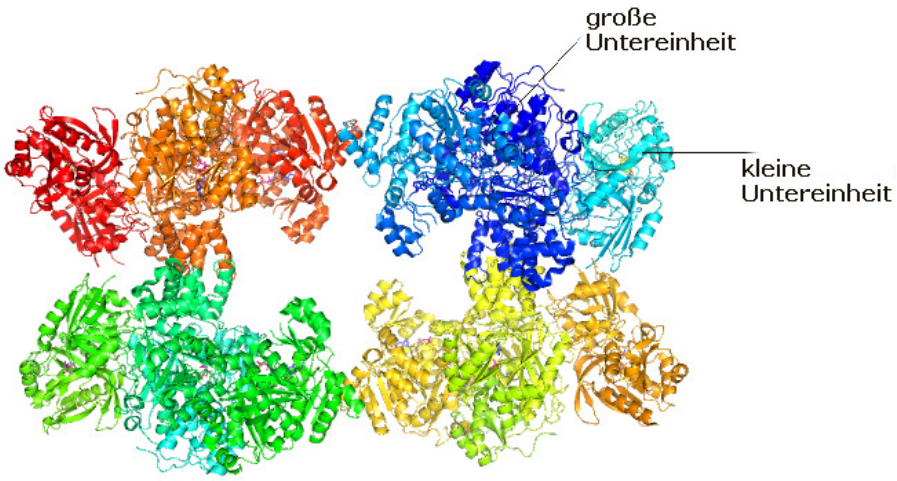


Abbildung 3-17: Carbamoylphosphat Synthase , PDB 1CS0, *Escherichia coli.*

Die vollständige Sequenz der Carbamoylphosphat Synthase variiert zwischen 197 Aminosäuren bei *Cricetulus griseus* bis über 2000 Aminosäuren zum Beispiel bei *Schizosaccharomyces pombe* und *Emericella nidulans*. Das Enzym bildet meist ein Dimer [140]. In *E. coli* besteht das Enzym aus 382 Aminosäuren (kleine Untereinheit) und 1073 Aminosäuren (große Untereinheit). Im vorliegenden Clusterbaum 31176 befinden sich die Sequenzen der großen Untereinheit.

Die meisten Publikationen beschränken sich auf die Analyse der Carbamoylphosphat Synthase aus *E. coli* und nur für diesen Organismus sind PDB-Einträge vorhanden. Die bisherigen und weiteren Aminosäurepositionsangaben dieses Beispiels beziehen sich somit auf die Sequenz von *Escherichia coli.*

3.4.3.2 Der katalytische Mechanismus von EC 6.3.5.5

Bei der Hydrolyse von Glutamin, die von der kleinen Untereinheit katalysiert wird, ist die katalytische Triade C269, H353 und E355 von großer Bedeutung. Dabei kann man die Rolle von C269 hervorheben. Diese Aminosäure ist die entscheidende Position während der Reaktion. Sie liegt in einer starken Drehung eines Loops, wie es in Enzymen üblich ist, die zur Familie der α/β-Hydrolasen gehören [160]. Sie initiiert eine nucleophile Attacke auf die Amidgruppe des Glutamins [159]. Die C269S Mutante bindet Glutamin, kann das Substrat aber nicht hydrolysieren [162]. Die Mutation H353D führt zu einem Anstieg des K_m-Werts.

Q273, H312 und S47 sind an der Bindung des Glutamins beteiligt. Während die Seitenkette von Q273 eine H-Bindung zur Carboxygruppe des Intermediates eingeht, stabilisiert S47 das Oxyanion während der Glutaminhydrolyse [162]. K202 und E355 sind an der Bildung des aktiven Zentrums beteiligt. Die Vermutung, dass die Aminosäure E355 bei der Katalyse eine entscheidende Rolle spielt, hat sich nicht bestätigt [163]. Die Aminosäuren G241, G243, L270, N311, G313 und F314 stabilisieren zudem über H-Brücken das Substrat bzw. Intermediate [162].

In der Großen Untereinheit sind zwei aktive Zentren vorhanden, die während der Reaktion ATP binden. Unter der Umsetzung von ATP zu ADP werden die Phosphorylierungen von Bicarbonat zu Carboxyphosphat und Carbamat zu Carbamoylphosphat katalysiert [111]. Es existieren zudem zwei intramolekulare Tunnel, die beide aktive Zentren der großen Untereinheit und das aktive Zentrum der kleinen Untereinheit miteinander verbinden. Der Ammoniak-Tunnel erstreckt sich von dem aktiven Zentrum der kleinen Untereinheit bis zum aktiven Zentrum der Carboxyphosphat Domäne. Auf diese Weise hat das Ammoniakmolekül keinen Kontakt zum Lösungsmittel [111]. Der Carbamat-Tunnel verbindet beide aktive Zentren der großen Untereinheit. Ammoniak gelangt nicht durch den Ammoniak-Tunnel zur großen Untereinheit, solange das Carboxyphosphat Intermediat nicht gebildet wurde. Auf diese Weise wird die Aktivität des Enzyms durch die Bildung des Carboxy Intermediats getriggert [112]. Dies wird durch Konformationsänderungen innerhalb des Enzyms, falls sich Intermediate gebildet haben, sowie durch allosterische Regulation des Enzyms, erreicht.

Die große Untereinheit lässt sich nach Thoden *et al.* [110] in die vier folgenden strukturellen Untereinheiten einteilen: Carboxyphosphat Synthetic Component, Oligomerization Domäne, Carbamoylphosphat Synthetic Component und Allosterische Domäne.

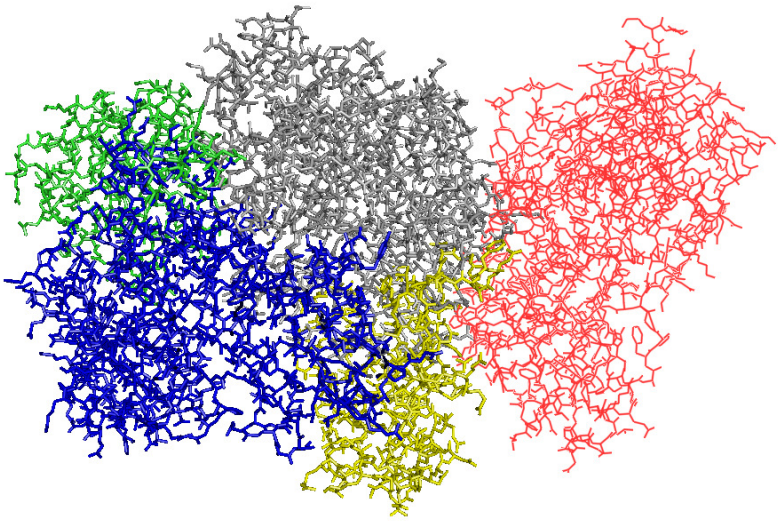


Abbildung 3-18: Carbamoylphosphat Synthase, PDB 1CS0, kleine Domäne und große Domäne aus *E. coli*. Die strukturellen Untereinheiten: Carboxyphosphat Synthetic Component (M1 bis E403) grau, Oligomerization Domäne (V404 bis A553) gelb, Carbamoylphosphat Synthetic Component (N554 bis N936) blau, Allosterische Domäne (S937 bis K1073) grün, kleine Untereinheit rot, nach Thoden *et al.* [110].

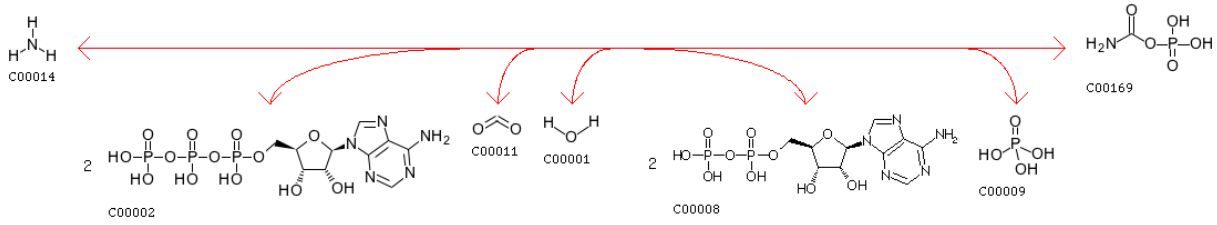
3.4.3.3 Weitere EC-Nummern im Clusterbaum und deren katalysierte Reaktionen

EC-Nummer 6.3.4.16:

Name des Enzyms: Carbamoylphosphat Synthase (ammonia)

Das Enzym Carbamoylphosphat Synthase katalysiert die Bildung von Kohlenstoff-Stickstoff Bindungen. Aus Ammoniak, Kohlendioxid und Wasser entstehen die Moleküle Orthophosphat und Carbamoylphosphat. Zwei Moleküle ATP werden zu zwei Moleküle ADP umgesetzt.

Reaktion: 2 ATP + NH_3 + CO_2 + H_2O ↔ 2 ADP + Orthophosphat + Carbamoylphosphat

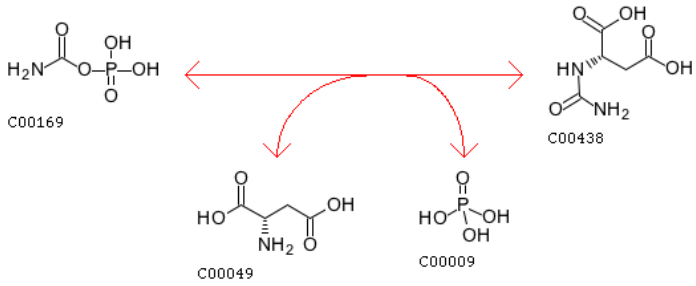


EC-Nummer 2.1.3.2:

Name des Enzyms: Aspartat Carbamoyltransferase

Das Enzym Aspartat Carbamoyltransferase katalysiert den Transfer einer Carbamoylgruppe von Carbamoylphosphat auf L-Aspartat. Dabei entstehen die Moleküle Orthophosphat und N-Carbamoyl-L-Aspartat.

Reaktion: Carbamoylphosphat + L-Aspartat ↔ Orthophosphat + N-Carbamoyl-L-Aspartat



3.4.3.4 Größte gemeinsame Teilstrukturen der Edukte, Produkte und Co-Substrate der Enzymreaktionen

Die größten gemeinsamen Teilstrukturen der Edukte, Produkte und Co-Substrate werden für die Reaktionen der EC-Nummern EC 6.3.5.5 + EC 6.3.4.16, EC 6.3.5.5 + EC 2.1.3.2 und EC 6.3.4.16 + EC 2.1.3.2 verglichen. Aufgrund der Vielzahl der an den Reaktionen beteiligten Moleküle und den damit verbundenen Kombinationsmöglichkeiten, beschränkt sich die Darstellung der Ergebnisse in den Tabellen auf die Vergleiche, bei denen die fünf größten gemeinsamen Teilstrukturen ermittelt wurden.

Vergleich EC 6.3.5.5 EC 6.3.4.16	**Molfile**	**Atome**	**Molekül**
	7261.mol	31	ATP
	7261.mol	31	ATP
c-MCS		31	
	6504.mol	27	ADP
	6504.mol	27	ADP
c-MCS		27	
	15530.mol	5	Phosphat
	15530.mol	5	Phosphat
c-MCS		5	
	8217.mol	8	Carbamoylphosphat
	8225.mol	8	Carbamoylphosphat
c-MCS		8	

	11282.mol	1	H_2O
	11282.mol	1	H_2O
c-MCS		1	

Tabelle 3-29: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 6.3.5.5 und EC 6.3.4.16.

Vergleich EC 6.3.5.5 EC 2.1.3.2	**Molfile**	**Atome**	**Molekül**
	8217.mol	8	Carbamoylphosphat
	8217.mol	8	Carbamoylphosphat
c-MCS		8	
	15530.mol	5	Phosphat
	15530.mol	5	Phosphat
c-MCS		5	
	8217.mol	8	Carbamoylphosphat
	15530.mol	5	Phosphat
c-MCS		5	
	12211.mol	10	L-Glutamat
	13821.mol	12	N-Carbamoyl-L-Aspartat
c-MCS		7	
	15530.mol	5	Phosphat
	8217.mol	8	Carbamoylphosphat
c-MCS		5	

Tabelle 3-30: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 6.3.5.5 und EC 2.1.3.2.

Vergleich EC 6.3.4.16 EC 2.1.3.2	**Molfile**	**Atome**	**Molekül**
	8225.mol	8	Carbamoylphosphat
	8217.mol	8	Carbamoylphosphat
c-MCS		8	
	15530.mol	5	Phosphat
	15530.mol	5	Phosphat
c-MCS		5	
	8225.mol	8	Carbamoylphosphat
	15530.mol	5	Phosphat
c-MCS		5	
	15530.mol	5	Phosphat
	8217.mol	8	Carbamoylphosphat
c-MCS		5	
	6504.mol	27	ADP
	15530.mol	5	Phosphat
c-MCS		5	

Tabelle 3-31: Vergleich der Moleküle aus den Reaktionen der EC-Nummern 6.3.4.16 und EC 2.1.3.2.

Wie den Tabellen 3-29 bis 3-31 entnommen werden kann, sind die Edukte, Produkte oder Co-Substrate der untersuchten Enzymreaktionen identisch oder sehr ähnlich. Mindestens ein Molekül ist bei jedem Vergleich identisch.

3.4.3.5 Vergleich der EC-Kombinationen mit der Clusterung nach identischen R-Strings

Die Enzyme mit den EC Nummern EC 6.3.5.5, EC 6.3.4.16 und EC 2.1.3.2 wurden aufgrund identischer R-Strings nicht gruppiert.

3.4.3.6 Untersuchung von Falsch-Positiven Treffern ausgewählter Sequenzmuster

Das Muster für EC 6.3.5.5, Clusterbaum 31176, E-Wert 10^{-84}. Das Muster besteht aus 210 Positionen und wurde aus 24 Sequenzen generiert. Bei der Suche gegen SWISS-PROT und TrEMBL wurden 43 Richtig-Positive und 19 Falsch-Positive Treffer erzielt. Alle Falsch-Positiven Treffer sind das Ergebnis aus der Suche gegen TrEMBL. Bei der Suche gegen SWISS-PROT ergaben sich keine Falsch-Positiven Treffer.

Die folgende Tabelle gibt einen Überblick über alle Falsch-Positiven Treffer des Musters für EC 6.3.5.5, Clusterbaum 31176, E-Wert 10^{-84}.

Sequenz	Start	Ende	Datenbank	Bezeichnung
Q1H390_METFK	7	672	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q3RDU0_XYLFA	7	678	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q1Y6W6_STAAU	7	670	TREMBL	Carbamoyl-phosphate synthase, large subunit
A1GKH2_9THEM	7	665	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q3R6G9_XYLFA	7	678	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q0ESF3_THEET	7	670	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q1QCJ1_PSYCK	35	708	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q3RBA9_XYLFA	7	678	TREMBL	Carbamoyl-phosphate synthase, large subunit
A1FYJ0_XANMA	7	678	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q1Y0D1_STAAU	7	670	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q8GGJ2_LACPL	7	668	TREMBL	CarB
A0LC16_MAGSM	7	676	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q1WXC1_9FIRM	7	672	TREMBL	Carbamoyl-phosphate synthase, large subunit
A0V6B1_COMAC	7	679	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q31H98_THICR	7	680	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q47HI6_DECAR	7	672	TREMBL	Carbamoyl-phosphate synthase, large subunit
Q5GYT0_XANOR	17	688	TREMBL	Carbamoyl-phosphate synthase large chain
Q4UU74_XANC8	7	678	TREMBL	Carbamoyl-phosphate synthase large chain
Q65JU7_BACLD	7	670	TREMBL	PyrAB (Carbamoyl-phosphate synthetase) (Catalytic subunit)

Tabelle 3-32: Untersuchung Falsch-Positiver Treffer des Sequenzmusters für EC 6.3.5.5 aus Clusterbaum 31176, E-Wert 10^{-84}.

Aus den Bezeichnungen der getroffenen Sequenzen (vgl. Tabelle 3-32) kann geschlossen werden, dass zumeist Sequenzen des Enzyms Carbamoylphosphat Synthase, große Untereinheit, getroffen werden. Diese wurde als Falsch-Positiv bewertet, da keine EC-Nummern angegeben wurden. Eine Übersicht der Kriterien, die für Beurteilung von Treffern maßgebend sind, liefert der Abschnitt 2.9 im Teil Daten, Algorithmen und Methoden.

3.4.3.7 PROSITE- und Clustermuster

Die im Clusterbaum 31176, E-Wert 10^{-84} enthaltenen Domänen erzielen den PROSITE Hit PS50975. Dieser Hit wird durch PDOC50975 dokumentiert. Demnach werden die im Cluster enthaltenen Domänen in die ATP-Grasp Superfamilie eingeordnet. Zu dieser Superfamilie gehören zur Zeit 17 Enzymgruppen an, die eine ATP abhängige Ligation von Carboxylat auf ein Molekül katalysieren, das eine Amino- oder Thiolgruppe enthält. Zu dieser Enzymgruppe gehören u.a. die Enzyme Urea Amidolyase, EC 6.3.4.6, Phosphoribosylamin Glycin Ligase EC 6.3.4.13, Biotin Carboxylase, EC 6.3.4.14, 5-(Carboxyamino) Imidazol Ribonucleotid Synthase, EC 6.3.4.18 an, die zur gleichen Subsubklasse gehören, wie die im Clusterbaum 31176 geclusterten Sequenzen der EC-Nummer 6.3.4.16. Das PROSITE-Profil hat eine Übereinstimmung (abhängig von der zur Suche verwendeten Sequenz) von etwa Aminosäure 130 bis etwa Aminosäure 330. Das PROSITE-Profil ist demnach wesentlich kürzer als die im Cluster enthaltenen Sequenzen, aus denen das Muster für EC 6.3.5.5 generiert wurde.

Neben dem oben beschriebenen PROSITE-Profil, sind zwei weitere Einträge für EC 6.3.5.5 in der PROSITE-Datenbank vorhanden. Beide Muster liegen im Bereich der großen Untereinheit der Carbamoylphosphat Synthase. Die Muster wurden von den Wissenschaftlern der PROSITE-Datenbank so gewählt, dass jeweils ein Muster im Bereich der aktiven Zentren der großen Untereinheit liegt. Die ausgewählten Aminosäuren spielen wahrscheinlich eine große Rolle bei der katalytischen Reaktion bzw. bei der Bindung von ATP. Die Muster PS00866 und PS00867, (vgl. Abbildungen 3-19 und 3-20) für die Carbamoylphosphat Synthase Untereinheit, werden in der Struktur PDB 1CS0, Abbildung 3-21, farbig markiert.

```
[FYV]-[PS]-[LIVMC]-[LIVMA]-[LIVM]-[KR]-[PSA]-[STA]-x(3)-[SG]-G-x-[AG]
```

Abbildung 3-19: PROSITE-Muster für EC 6.3.5.5, PS00866.

```
[LIVMF]-[LIMN]-E-[LIVMCA]-N-[PATLIVM]-[KR]-[LIVMSTAC]
```

Abbildung 3-20: PROSITE-Muster für EC 6.3.5.5, PS00867.

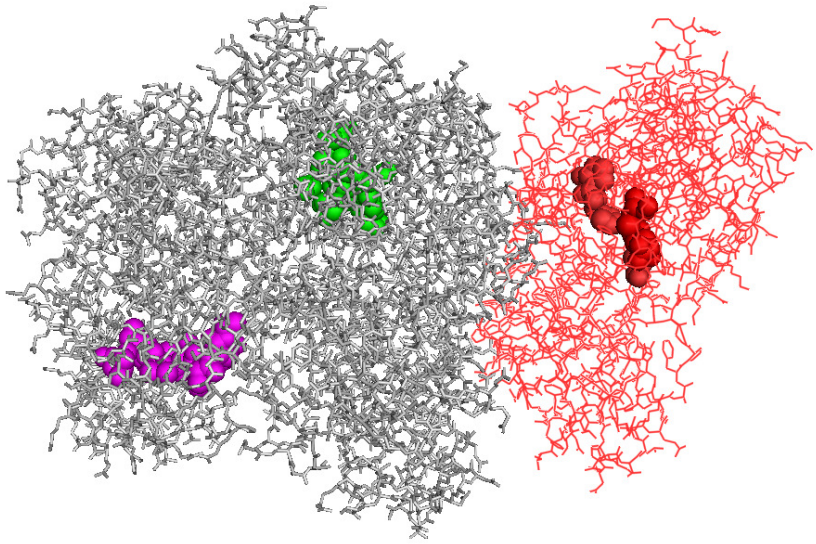


Abbildung 3-21: Carbamoylphosphat Synthase, PDB 1CS0, große Untereinheit grau, kleine Untereinheit rot, *E. coli*. Muster aus der PROSITE-Datenbank wurden in der großen Domäne pink (Y838 bis A845) und grün (F164 bis S171, G175, G176, G178) markiert. In diesen Bereichen liegen die aktiven Zentren. In der kleinen Untereinheit wurde das aktive Zentrum in rot (C269, H353, E355, S47, P354) hervorgehoben.

Das automatisch generierte Muster für EC 6.3.5.5, Clusterbaum 31176, E-Wert 10^{-84} besteht aus 210 Positionen und ist damit wesentlich länger als die Muster der PROSITE-Datenbank. In diesem Cluster sind 24 Sequenzen vorhanden, die ausschließlich zur EC-Nummer 6.3.5.5 gehören. Die farbigen Markierungen zeigen die Musterpositionen, die mit dem Muster PS00866 aus PROSITE identisch sind. Das Muster PS00867 liegt nicht im Bereich des Clustermusters.

```
[IL]-x(4)-[ILV]-[IL]-G-[AS]-G-x(3)-[IV]-x-[HQ]-x(2)-E-[FL]-D-x-[AS]-x(2)-Q-
x(4)-[FL]-[KR]-x(6)-[IV]-[IL]-x-[DN]-x-N-x(2)-[ST]-[IV]-x(2)-[DE]-x(2,4)-D-
x(11)-[ILV]-x(2)-I-[FIL]-x(4)-P-D-x-[IL]-[IL]-x(2)-[ILMV]-G-x-[QR]-x(2)-
[FL]-[EN]-x-[AT]-x(6)-G-[ILV]-[FIL]-x(5)-[EKQ]-[ILV]-[IL]-G-x(4)-[AST]-I-
x(3)-[EN]-x(4)-[FL]-x(3)-[LM]-x(28)-P-x-[IV]-[IV]-R-x(7)-[AST]-x(3)-[IL]-
x(6)-[FL]-x(7)-[FIL]-x(2)-S-x(5)-[LV]-[ILV]-[DEQ]-x(2)-[ILV]-x-G-[FWY]-x-E-
x-E-x(2)-V-x-R-D-x(5)-[ILMV]-x-[ILV]-x(2)-[IM]-E-[DN]-x-D-P-[IMV]-G-[IV]-H-
[AT]-G-[DE]-S-x(4)-P-x(2)-T-L-x-[DN]-x-[DEQ]-x(3)-[ILM]-[KR]-x(2)-[AS]-x(2)-
[IV]-x(3)-[ILV]-x-[IV]-x(4,5)-[HN]-[IV]-Q-x(10,12)-[IV]-[EK]-x(2)-P-x(3)-R-
x-[ST]-A-[FL]-x-S-[KQ]-[AS]-T-G-[FY]-P-I-A-[KQR]-[ILMV]-[AST]-[AS]-x-[IL]-
x(2)-G-x(2)-L-x-[DE]-[ILM]-x(9,10)-A-x-[FI]-E-P-[AST]-[ILM]-D-[HY]-x(3)-
[KR]-x-P-x-[FW]-x-[FL]-[DEN]-x-[FL]-x(6)-[IL]-x(5)-[AS]-x-G-x(2)-[FM]-x-
[IMV]-x(6)-[AS]-x(2)-K-x(21,39)-[ILV]-x(6)-R-[ILV]-x(2)-[ILV]-x(3)-[FILM]-
x(8)-[ILV]-x(5)-[IV]-x(2)-[FW]-[FY]-[FIL]-x(2)-[FILM]-x(18,30)-L-x(3)-K-
x(8,21)-[ILM]-x(10)-[ILV]-x(6,8)-[ILV]-x(3)-[FY]-K-x-[IV]-[DE]-x(4)-E-[FI]-
x(3)-[ST]-x(3)-Y-x-[AST]-x(5,6)-E-x(8,9)-[ILMV]-[IV]-[IL]-G-x-G-x(2)-R-[IL]-
G-x(5)-D-Y-x(9)-[HKQR]-x(7)-[IMV]-[IMV]-N-x-N-P-x-[ST]-V-[ST]-x-D-x(4)-[DE]-
[KR]-x-[FY]-x-E-P-x-[ST]-x-E-x-[ILV]-x(2)-[IV]-x(3)-E-x-P-x(2)-[ILV]-x-
[ILV]-x(4)-[DQ]-x(2)-[ILV]-[KN]-x(11,22)-[IV]-[IL]-x(11)-D-R-x(2)-[FY]
```

Abbildung 3-22: Muster für EC 6.3.5.5 aus Clusterbaum 31176. Positionen, die den Musterpositionen aus PROSITE entsprechen sind rot markiert.

3.4.3.8 Identifizierungen der generierten Muster in den 3D-Darstellungen der Enzyme

Die Einordnung des Musters aus Clusterbaum 31176, E-Wert 10^{-84} in die 3D-Struktur des Moleküls ist aufgrund seiner Länge von 210 Positionen nicht sinnvoll, da die große Anzahl der Musterpositionen kein klares Bild liefert. Stattdessen wird in der folgenden Abbildung das Muster aus dem besprochenen Clusterbaum 31176 dargestellt, das aus 311 Sequenzen bei E-Wert 10^{-2} generiert wurde. Dieses Muster besteht aus 75 Positionen und ist damit wesentlich kürzer im Vergleich zu dem oben dargestellten Muster. Das Muster wird in Abbildung 3-23 dargestellt.

```
[ILMV]-x(6)-[ILV]-x(8,15)-[FM]-x(25,99)-P-x-[ILMV]-x(35,58)-[ILV]-x(8)-[EQ]-
x-E-x(2)-[ILMV]-x(2)-D-x(10,11)-[ILM]-E-[DN]-x(6)-H-[AT]-G-[DE]-x(5)-P-x(3)-
[ILM]-x(7,8)-[KQR]-x(16,19)-[HN]-[IV]-Q-x(10,15)-[IV]-[EK]-x(6)-R-x-[ST]-A-
[FL]-x-S-[KQ]-x-[AST]-G-[FY]-x-[IL]-A-x(11)-L-x(2)-[ILMV]-x(12,21)-P-x(2)-D-
[HY]-x(3)-[KR]-x(14,22)-[ILMV]-x(5)-[AS]-x-G-x(2)-[FIM]-x-[ILMV]-x(9)-K-
x(54,99)-[FWY]-[FY]-x(31,54)-[ILMV]-x(18,56)-[IV]-x(5)-E-x(5,8)-Y-x(18,89)-
[ILV]-G-x-G-x(3)-[ILV]-x(6)-D-[WY]-x(17)-[ILMV]-x-N-x-N-P-x-[ST]-V-[ST]-x-D-
x(7)-[FY]-x(6)-[DE]-x-[ILV]-x(2)-[ILV]-x(21,32)-[FILM]-x(20,46)-R-x(2)-[FWY]
```

Abbildung 3-23: Muster für EC 6.3.5.5 aus Clusterbaum 31176, E-Wert 10^{-2}.

Die Musterpositionen des Musters für EC 6.3.5.5, E-Wert 10^{-2} aus Clusterbaum 31176 wurde in der Abbildung 3-24 farbig markiert.

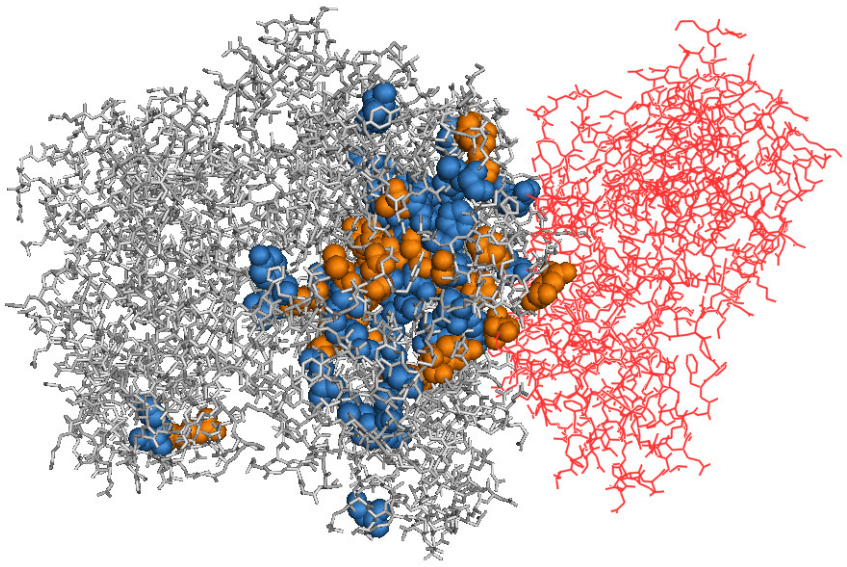


Abbildung 3-24: Carbamoylphosphat Synthase, PDB 1CS0, große Untereinheit grau, kleine Untereinheit rot, *E. coli*. Muster für EC 6.3.5.5 aus Clusterbaum 31176, E-Wert 10^{-2}. Konservierte Musterpositionen blau, hochkonservierte Positionen orange, Sequenzbereich des Musters: I116 bis F678.

Das generierte Muster liegt im Sequenzbereich der großen Untereinheit der Carbamoylphosphat Synthase zwischen den Aminosäuren I116 und F678. Bei einer Sequenzlänge von 1073 Aminosäuren der großen Untereinheit liegen die Aminosäuren des Sequenzmusters relativ zentral. Das Muster erstreckt sich damit nach der strukturellen Einteilung der großen Untereinheit in vier Strukturen nach Thoden *et al.* [110] über die Carboxyphosphat Synthetic Component, der Oligomerization Domäne und der Carbamoylphosphat Synthetic Component. Die große Untereinheit der Carbamoylphosphat Synthase besitzt zwei aktive Zentren, die während der Katalyse ATP binden (vgl. Abschnitt 3.4.3.2). Die Positionen des generierten Muster liegen nicht in der Nähe der aktiven Zentren, obwohl diese Bereiche typischerweise stark konserviert sind. Die Aminosäuren in diesen Bereichen wurden für die PROSITE-Muster ausgewählt, da sich diese Bereiche nach Meinung der Wissenschaftler, die die Muster in PROSITE erstellt haben, eignen, mit diesen Mustern Enzyme aus der gleichen Familie zu identifizieren. In Abbildung 3-21 wurden die Muster der PROSITE-Datenbank farbig markiert. Da im Vergleich zu Abbildung 3-24 die gleiche Perspektive gewählt wurde, lassen sich beide Moleküle gut vergleichen. Anhand der Markierungen der Musterpositionen innerhalb der 3D-Strukturen des Enzyms, lässt sich eine eindeutige Konzentrierung der Positionen auf einen Teilbereich der großen Untereinheit identifizieren. Die Aminosäuren des Musters liegen im Bereich, der in Kontakt mit der kleinen Untereinheit des Enzyms steht. Die kleine Untereinheit katalysiert die Hydrolyse von Glutamin zu Glutamat und Ammoniak. Das Reaktionsintermediat gelangt durch einen intramolekularen Tunnel von der kleinen Untereinheit zu der großen Untereinheit. Der Kontakt beider Untereinheiten ist daher essentiell für die Funktion des Enzyms. Eine starke Konservierung von Aminosäuren im Kontaktbereich der Untereinheiten, sowie im Bereich des intramolekularen Tunnels, wurde durch das Sequenzmuster nachgewiesen.

3.4.3.9 Vergleiche der erhaltenen Sequenzdomänen für EC 6.3.5.5 mit anderen Datenbanken

Wie der Tabelle 3-32 zu entnehmen ist, erstreckt sich die Proteindomäne für die Sequenzen des Clusterbaums 31176, E-Wert 10^{-84}, EC 6.3.5.5, auf die Sequenzabschnitte von etwa Aminosäure 7 bis Aminosäure 675.

Einträge für EC 6.1.5.5 für die Proteindomäne AS 7 bis AS 675 in anderen Datenbanken:

CATH:

Einordnung der Domäne in die homologe Superfamilie (3.30.470.20).
ATP-grasp fold, B domain.

Einordnung der Domäne in die Homologe Superfamilie (3.40.50.20).
[PDB] Transferase.

Pfam:

Einordnung der Domäne in die Familie CPSase_L_D2 (PF02786).

- Carbamoyl-phosphate synthase L chain, ATP binding domain.
- Die Domänenregion variiert zwischen AS 116 bis AS 417.

Einordnung der Domäne in die Familie CPSase_L_D3 (PF02787).

- Carbamoyl-phosphate synthetase large chain, oligomerisation domain.
- Die Domänenregion variiert zwischen AS 128 bis AS 417.

Einordnung der Domäne in die Familie CPSase_L_chain (PF00289).

- Carbamoyl-phosphate synthase L chain, N-terminal domain.
- Die Domänenregion variiert zwischen AS 116 bis AS 475.

4 Diskussion und Ausblick

4.1 Diskussion

4.1.1 Clusterung der Sequenzen

Die Clusterung von Daten ist in der Biologie eine enorm wichtige Methode, um große Datenmengen zu analysieren und um zusätzliche, wichtige Informationen wie eine Gruppierung, Selektion oder Klassifikation von Sequenzen eines Datensatzes zu erhalten. Diese Informationen, die üblicherweise automatisch mittels Algorithmen gewonnen werden, gehen über die simplen Informationen von Rohdaten hinaus. Datensätze von Proteinen und ganzen Genomen sind für eine Clusteranalyse besonders gut geeignet, da durch verschiedene Sequenzierungsprojekte auf der ganzen Welt eine Datenmenge erreicht wird, die manuell nicht in einer adäquaten Zeit analysiert werden kann. Ein Rückgang der Wachstumsgeschwindigkeit von Sequenzdatenbanken ist nicht abzusehen. Im Gegenteil ist es wahrscheinlich, dass durch verbesserte Sequenzierungstechniken die Geschwindigkeit, in der Sequenzdatenbanken wachsen, weiter zunimmt. All diese Tatsachen führen zu der Motivation, die Technik einer Clusterung von Proteinsequenzen zu perfektionieren. Dabei ist das Prinzip der Clusterung relativ simpel. Elemente werden so gruppiert, dass die Ähnlichkeit innerhalb einer Gruppe relativ groß, die Ähnlichkeit zwischen den Gruppen aber relativ klein ist [75, 76].

Dieses Prinzip der automatischen Analyse von Proteinsequenzen wurde bereits von mehreren Forschergruppen umgesetzt. So hat Krause *et al.* eine Clusterung von mehr als einer Million Sequenzen durchgeführt und auf diese Weise die erhaltene Gruppierung genutzt, um Proteinfamilien und Proteinsuperfamilien zu bilden [77]. Yona *et al.* entwickelte ProtoMap, eine automatische Klassifizierung aller Proteine der SWISS-PROT Datenbank. Dieser Ansatz basiert auf der paarweisen Analyse von Proteinsequenzen und einer anschließenden Clusterung der Ergebnisse [78]. Enright und Ouzounis nutzen die Idee der automatischen Clusterung von Sequenzen, um große Datensätze, wie z.B. komplette Genome, zu analysieren. Ihre Einteilung der Sequenzen in Familien basiert auf Sequenzähnlichkeit [79].

In dieser Arbeit wurde die Clusterung von Proteinsequenzen aus den Datenbanken SWISS-PROT und TrEMBL dazu genutzt, um eine biologisch sinnvolle Klassifizierung von Enzymsequenzen zu erhalten. Dabei wurden die Daten nach der etablierten Methode des *single linkage* Verfahrens geclustert. Diese Form der Clusterung basiert auf der Idee,

dass zu Anfang der Clusterung Daten, die sehr ähnlich zueinander sind, geclustert werden. Mit zunehmender Distanz werden die Daten immer stärker geclustert, so dass ein Clusterbaum entsteht. Bezogen auf Proteinsequenzen, sind alle Sequenzen innerhalb eines Clusterbaums ähnlich, d.h. homolog zueinander.

Aus den Datenbanken SWISS-PROT und TrEMBL wurden 239923 Enzymsequenzen extrahiert, die mindestens eine vollständige EC-Nummer tragen. Diese Sequenzen wurden durch lokale BLAST all-vs-all Alignments in Domänen eingeteilt. Dadurch sind 2108181 Subsequenzen entstanden, die im Verlauf der Analyse dieser Daten geclustert wurden. In Diagramm 3-1 ist eine Übersicht über die Clusterung der Sequenzen dieser Arbeit dargestellt. Bei E-Wert 10^{-2} existieren 38691 Cluster, bei E-Wert 10^{-181} sind 1971956 Cluster vorhanden. Es sind somit 38691 Clusterbäume entstanden, also Gruppierungen von Sequenzen, deren Mitglieder untereinander homolog sind.

Die Beurteilung der Signifikanz eines Grenzwertes ist schwierig. Der größte in dieser Arbeit verwendete Grenzwert bei der Clusterung ist der E-Wert 10^{-2}. Dieser Wert wurde genutzt, da verschiedene Arbeiten gezeigt haben, dass dies der größte E-Wert ist, bei dem mit einer großen Sicherheit davon ausgegangen werden kann, dass eine Sequenzähnlichkeit nicht zufällig detektiert wurde [80]. Allerdings wurde bei E-Wert 10^{-2} eine Vielzahl an Sequenzen geclustert, so dass ein Cluster mit 991205 Sequenzen entstanden ist. Die Ermittlung der Ursache für diese Clusterung ist schwierig und kann nicht eindeutig geklärt werden. Es ist aber fraglich, ob eine Clusterung dieses Ausmaßes biologisch signifikant ist. Möglicherweise wäre es sinnvoll, den Grenzwert auf 10^{-3} zu senken, um Cluster zu vermeiden, die annähernd eine Million Sequenzen enthalten. Mit zunehmender Größe von SWISS-PROT und TrEMBL, steigt auch die Gefahr, dass Cluster bei E-Wert 10^{-2} entstehen, die über eine Million Sequenzen hinausgehen.

4.1.2 Ermittlung der Domänenstruktur

Proteine bestehen aus einer oder mehreren Proteindomänen (vgl. Abschnitt 1.2). Um diese Untereinheiten zu identifizieren, wurden Methoden entwickelt, die Strukturdaten der Proteine nutzen. Die Datenbanken CATH [81] und SCOP [82] sind auf diese Weise entstanden. Eine weitere etablierte Datenbank ist ProDom [58]. Diese Datenbank bezieht ihre Informationen aus PSI-BLAST Alignments. Die Datenbank Pfam beruht auf

Ergebnissen von erstellten *Hidden Markov Models* und anschließenden multiplen Sequenzalignments [59].

In dieser Arbeit wurden paarweise BLAST all-vs-all Alignments aller in der Datenbank enthaltenen Sequenzen durchgeführt. Wurden besonders stark konservierte Sequenzabschnitte identifiziert, wurde daraus geschlossen, dass sich diese Abschnitte aufgrund einer besonderen Funktion im Verlauf der Evolution erhalten haben. Dabei wurde aber nicht nur ein einzelner Sequenzabschnitt pro Sequenz als hoch konserviert identifiziert, denn es kam auch vor, dass verschiedene Abschnitte einer Sequenz zu verschiedene Abschnitten anderer Sequenzen stark ähnlich sind. Alle auf diese Weise erhaltenen Subsequenzen werden in die Datenbank eingetragen. Einen Hinweis, welche biologische Signifikanz eine bestimmte Domäne hat, liefert der E-Wert, der bei jedem paarweisen Vergleich gespeichert wird. Eine weitere Einschätzung über die Bedeutung eines bestimmten Sequenzabschnitts liefert neben der experimentellen Bestimmung im Labor ein Vergleich der erhaltenen Ergebnisse mit den Ergebnissen anderer Methoden, bzw. daraus erstellten Datenbanken.

Im Ergebnisteil wurden drei Beispiele behandelt, in denen u.a. die erhaltenen Domänen mit den Ergebnissen anderer Datenbanken exemplarisch verglichen wurden. Im ersten Beispiel, EC 3.5.99.6 sind im Clusterbaum 19905 die Domänen von Aminosäure (AS) 66 bis Aminosäure (AS) 230 und von Aminosäure (AS) 1 bis Aminosäure (AS) 240 vorhanden. Die erste Domäne liegt damit innerhalb des Sequenzbereichs der zweiten Domäne. Dieser Sequenzabschnitt wird in der Datenbank CATH in die homologe Superfamilie 3.40.50.1360 eingeordnet und weist der Domäne eine Hydrolasefunktion zu. Die Datenbank Pfam ordnet den erhaltenen Sequenzabschnitt in die Familie Glucosamine_iso (PF01182) ein. Die Domäne der Glucosamine-6-phosphate isomerase/ 6-phosphogluconolactonase liegt nach Pfam im Bereich von AS 8 bis AS 230. Die Datenbank SCOP [107] gruppiert den Bereich von AS 1 bis AS 240 in die Klasse Alpha/Beta proteins (a/b), Fold NagB/RpiA/CoA transferase-like, ein.

Im zweiten Beispiel wurde im Clusterbaum 18685, E-Wert 10^{-33}, für das Enzym EC 5.1.3.4 die Proteindomäne von AS 56 bis AS 202 ermittelt. Für das Enzym EC 4.1.2.17, E-Wert 10^{-14}, wurde im gleichen Clusterbaum die Sequenzgrenzen von AS 10 bis AS 100 bestimmt. Der Sequenzabschnitt für EC 5.1.3.4 wird in der Datenbank SCOP in die Klasse Alpha/Beta proteins (a/b) eingeordnet, die Datenbank PANTHER gruppiert den Sequenzabschnitt in die Familie FUCULOSE PHOSPHATE ALDOLASE (PTHR22789)

ein. In der Datenbank Pfam entspricht der Sequenzabschnitt für EC 4.1.2.17, wie auch der Sequenzabschnitt für EC 5.1.3.4, der Familie Aldolase_II and Adducin N-terminal domain (PF00596). Ebenso existiert in CATH für beide Sequenzabschnitte der Enzyme EC 5.1.3.4 und EC 4.1.2.17 nur ein Eintrag. CATH gruppiert diese Sequenzabschnitte in die homologe Superfamilie L-Fuculose-1-Phosphat Aldolase ein.

Im dritten Beispiel wurde die Proteindomäne für EC 6.1.5.5, Clusterbaum 31176, E-Wert 10^{-84} bestimmt. Der Abschnitt liegt im Bereich von AS 7 bis AS 675. Die Datenbank CATH gruppiert diesen Bereich in die homologen Superfamilien ATP-grasp fold, B domain und [PDB] Transferase ein. In Pfam existieren drei Einträge für die Carbamoyl-phosphat Synthase, die im Bereich von AS 116 bis AS 417, AS 128 bis AS 417 und AS 116 bis AS 475 liegen.

Wie aus den Beispielen ersichtlich ist, werden die erkannten Domänen auch in anderen Datenbanken als katalytisch aktive Domänen erkannt. Nur im letzten Beispiel erstreckt sich die mittels Alignments und Clusteranalyse ermittelte Domäne über einen Großteil der Sequenz. Da der E-Wert mit 10^{-84} relativ klein und damit besonders verlässlich ist, existieren sehr viele Sequenzen, die über die komplette Länge besonders konserviert sind. Es liegt daher die Vermutung nahe, dass dieses Enzym aus nur einer Domäne besteht. Ein Vergleich mit anderen Datenbanken zeigt, dass das Enzym in mehrere katalytische Domänen eingeteilt werden kann. Das Ergebnis der Domänenstrukturvorhersage mit Hilfe von Alignments und Clusterung ist im Fall des letzten Beispiels nicht optimal. Dennoch ist das Ergebnis mit den ermittelten Sequenzalignments nachvollziehbar. Die Einschätzung des Clusterergebnisses in Betracht aller Beispiele ist insgesamt als positiv zu bewerten. Mit Hilfe der Clusteranalyse konnte eine Klassifizierung von homologen Sequenzen erhalten werden, die eine zufriedenstellende Einteilung von homologen Enzymdomänen liefert.

4.1.3 Bestimmung von Sequenzmustern

Sequenzmuster geben die Anordnung von bestimmten konservierten Aminosäuren wieder, die für die jeweilige Sequenz bzw. Proteinfamilie typisch sind. Dabei eignen sich die Sequenzen von Enzymen hervorragend für die Erstellung von Mustern, da sich diese im Laufe der Evolution weitgehend konserviert haben [2]. Besonders Aminosäuren, die bei der Katalyse, bei der Bindung von Liganden und bei der Etablierung einer 3D-Struktur beteiligt sind, wurden während der Evolution besonders stark konserviert und daher sind

meist diese Positionen in Sequenzmustern enthalten [54]. Aminosäuren des aktiven Zentrums von Enzymen sind immer funktionell konserviert [68].

Mehrere Forschergruppen entwickelten Algorithmen für die Identifizierung und Speicherung von Sequenzmustern. Sobel und Martinez entwickelten ein Programm zur Erstellung von multiplen Sequenzalignments [13]. Während der Durchführung dieser Alignments wurden Substrings bestimmt, die eine Minimallänge überschreiten mussten. Da für die Erstellung eines Alignments mit diesen Substrings in den Sequenzen des multiplen Alignments gesucht wurde, kann dieser Ansatz als erstmalige Verwendung von Sequenzmustern gewertet werden. Der MOTIF Algorithmus wurde von Smith *et al.* entwickelt [36]. Nach ihrer Idee werden konservierte Blöcke von einer Länge von drei Aminosäuren identifiziert. Zwischen diesen Blöcken sind Lücken erlaubt. Das Programm PRATT von Jonassen *et al.* [22] führt Muster mit variablen Positionen und mit flexiblen Lücken ein. Es werden mehrere Sequenzmuster unterschiedlicher Qualität erzeugt, abhängig von den jeweils genutzten Einstellungen.

Die größte, bekannteste und öffentlich zugängliche Datenbank, die Sequenzmuster generiert und speichert ist PROSITE [56] (vgl. Abschnitt 2.11.3). Die Muster werden manuell ausgewählt und erstellt. Die meisten dieser Muster beschreiben Aminosäuren des aktiven Zentrums [115]. Zusätzlich werden automatisch erstellte Profile zur Verfügung gestellt.

Alle dargestellten Algorithmen haben ihre Vor- und Nachteile. Werden Muster automatisch erstellt, besteht immer die Gefahr, dass wichtige Aminosäuren in einem Muster nicht berücksichtigt werden. Abhängig vom verwendeten Algorithmus bestehen Limitierungen in Sequenzanzahl oder Musterlänge. Die dargestellten Ansätze sind teils unflexibel, da keine Lücken erlaubt sind oder es bestehen Beschränkungen der Sequenzlängen. Manuelle Ansätze zur Identifizierung von konservierten Positionen sind in der Regel sehr genau. Da durch verschiedene Sequenzierungsprojekte und durch die Verbesserung von Sequenzierungsmethoden Sequenzdatenbanken sehr schnell wachsen, ist es sehr schwierig, die steigende Sequenzanzahl manuell zu bewältigen.

In dieser Arbeit wurde das Programm CLUSTAL W genutzt, Alignments von Sequenzen zu erstellen. CLUSTAL W markiert in einem Alignment konservierte Aminosäuren. Aus diesen Markierungen werden Muster im PROSITE-Format erstellt.

Dieser Ansatz wurde gewählt, da auf diese Weise keine Limitierungen in Sequenzlänge oder –anzahl bzw. Musterlänge existiert. Zudem entsteht während der Erstellung des Musters ein Alignment, mit dessen Hilfe es möglich ist, die Bestimmung der konservierten Positionen nachzuvollziehen.

Es existiert eine Vielzahl von Programmen, die globale Alignments durchführen. Als Beispiele seien z.B. die Programme AMAP [16], MAVID [17] und MUSCLE [14, 15] genannt. Die Entscheidung, bei dieser Arbeit CLUSTAL W zu verwenden und nicht die oben aufgeführten Programme, beruht auf dem Umstand, dass CLUSTAL W das populärste dieser Programme ist und sich die Alignments, die Testweise durchgeführt wurden, sich in der Rechengeschwindigkeit, sowie im Resultat nicht wesentlich von der Rechengeschwindigkeit und vom Resultat von CLUSTAL W unterscheiden. Im Vergleich mit verschiedenen Programmen erstellt CLUSTAL W tendenziell die besten Alignments [116].

Die Sequenzmuster werden im PROSITE-Format ausgegeben. Es ist das bekannteste Format, das von der bekanntesten Sequenzmusterdatenbank entwickelt wurde. Die Syntax dieses Formats ist für den Menschen leicht zu lesen und die automatisch generierten Muster aus dieser Arbeit können direkt mit Mustern der PROSITE-Datenbank verglichen werden. Es existieren weitere Formate zur Speicherung von Sequenzmustern. Diese stellen meist eine Abwandlung des PROSITE-Formats dar (vgl. Abschnitt 2.8.1).

4.1.4 Die Zusammenfassung der Sequenzen

Da Cluster teilweise sehr viele Sequenzen enthalten, wurde für die Durchführung von Alignments die Sequenzanzahl reduziert, indem Sequenzregionen, die vollständig mit anderen Regionen der gleichen Sequenz überlappen, aus dem Alignment ausgeschlossen wurden (vgl. Abschnitt 2.6). Dieser Vorgang wurde automatisiert mit Teilen des Programms *os_sys.py* durchgeführt. Auf diese Weise wurden Alignments möglich, die aufgrund ihrer hohen Sequenzzahl zeitlich nicht realisierbar wären.

Dieser Schritt war nötig, da BLAST Regionen identifiziert, die sich nur um wenige Aminosäuren am Anfang und Ende der Sequenz unterscheiden. Der Informationsgehalt zusammengeführter Sequenzen geht nicht verloren. Die Angaben der Sequenzzahlen in den Clusterbäumen sind aufgrund der Methode zur Identifizierung von Domänen mittels BLAST etwas irreführend, da zum Beispiel hundert Sequenzen in einem Cluster aufgrund

ähnlicher Regionen, möglicherweise achtzig zusammengeführter Sequenzen entsprechen. Die Sequenzanzahl sehr großer Cluster konnte durch die Zusammenfassung um bis zu 30% ihrer ursprünglichen Sequenzzahl ohne Informationsverlust reduziert werden. Je mehr Sequenzen ein Cluster enthält, bzw. je größer der E-Wert wird, desto größer ist die Wahrscheinlichkeit, dass überlappende Sequenzregionen in einem Cluster vorhanden sind.

4.1.5 Alignments mit CLUSTAL W

Die Anzahl der Sequenzen, auf deren Grundlage Alignments durchgeführt wurden, variiert stark. Meist waren pro Datensatz sehr wenige, bis etwa 20 Sequenzen für ein Alignment vorhanden, aber auch Alignments aus bis zu 1000 Sequenzen wurden durchgeführt (vgl. Diagramm 3-10). Obwohl CLUSTAL W Alignments mit hoher Geschwindigkeit erstellt, eignet sich das Programm optimal, wenn wenige Sequenzen im Datensatz vorhanden sind. Nicht nur der enorme Zeitaufwand von Alignments mit sehr vielen Sequenzen spielt dabei eine Rolle, vielmehr kann CLUSTAL W aus sehr vielen Sequenzen kein optimales Alignment erstellen, so dass kaum oder keine konservierten Positionen identifiziert werden, falls sehr viele Sequenzen im Alignment vorhanden sind. Das Unvermögen von CLUSTAL W, konservierte Positionen im Alignment zu identifizieren, liegt auch an der Unähnlichkeit der Sequenzen, da Cluster mit sehr hohen Sequenzanzahlen erst bei hohen E-Werten entstehen.

CLUSTAL W ist ein Programm, dessen Parameter für die Erstellung eines Alignments individuell angepasst werden können. Abhängig von der Beschaffenheit des Sequenzdatensatzes, können aufgrund möglicher bekannter Sequenzähnlichkeit oder aufgrund unterschiedlicher Sequenzlängen, CLUSTAL W Einstellungen individuell gewählt werden, z.B. die Definition der maximalen Anzahl von Lücken oder die Wahl der verwendeten Ähnlichkeitsmatrix. Da annähernd 120000 Alignments durchgeführt wurden, konnte CLUSTAL W aufgrund dieser großen Anzahl nicht für jedes einzelne Alignment manuell eingestellt werden. Stattdessen wurden alle Standardeinstellungen übernommen. Es war daher zu erwarten, dass in einigen Fällen nicht ein optimales Alignment gefunden wird und möglicherweise wichtige Positionen im Sequenzmuster fehlen. Auf die Qualität eines Alignments haben die Faktoren Sequenzanzahl im Alignment, die Länge dieser Sequenzen und die Ähnlichkeit der Sequenzen untereinander, Einfluss.

Für die Extraktion der Muster wurde ein Python Script verwendet, das während dieser Arbeit entwickelt wurde. Es liest CLUSTAL W Alignments ein und identifiziert markierte (hoch-)konservierte Positionen. Dieses Verfahren hat sich während der Arbeit bewährt, da die Extraktion einfach, schnell und automatisiert durchgeführt werden konnte, unabhängig von der Sequenzanzahl und Länge der verwendeten Sequenzen. Die automatisierte Eintragung der Muster in die Datenbank wurde größtenteils manuell überwacht und im Verlauf der Arbeit immer wieder optimiert. Aufgrund des Funktionsprinzips von CLUSTAL W unterscheiden sich die Sequenzmuster der Clustersequenzen von Mustern der PROSITE-Datenbank in einem Punkt: CLUSTAL W markiert in den erstellten Alignments identische oder funktionell konservierte Aminosäuren. So ist z.B. [FILM] eine funktionell konservierte Position, wie sie CLUSTAL W markieren könnte. Die PROSITE-Datenbank geht einen Schritt weiter und fasst an einer funktionell konservierten Position oft alle Aminosäuren zusammen, die an dieser Position in verschiedenen Proteinen vorkommen können. In Abschnitt 3, Abbildung 3-20 ist z.B. die Musterposition [LIVMSTAC] vorhanden. Dies wäre wie erwähnt bei CLUSTAL W nicht möglich. CLUSTAL W würde diese Position nicht in dieser Form als konserviert erkennen, da sich die Aminosäuren dieser Position zu stark voneinander unterscheiden.

4.1.6 Vergleich der erstellten Muster mit Mustern der PROSITE-Datenbank

Ein Vergleich der extrahierten Clustermuster mit den Mustern der PROSITE-Datenbank zeigte, dass die PROSITE-Datenbank Muster enthält, die sich nicht über sämtliche konservierte Aminosäuren einer Proteinsequenz erstrecken. Stattdessen wird für das Muster oft eine isolierte, konservierte Region aus der Proteinsequenz ausgewählt, die katalytisch wichtige Aminosäuren enthält. Dieses Muster soll Proteine, die dieses Muster enthalten, identifizieren. Da die PROSITE-Muster nicht alle konservierten Positionen der gesamten Sequenz bzw. Domäne enthalten, im Gegensatz der Muster, die mit Hilfe von CLUSTAL W erstellt wurden, sind die Muster der PROSITE-Datenbank kürzer, als die automatisch generierten Muster, die aufgrund von Alignments entstanden. CLUSTAL W erkennt im optimalen Fall alle konservierten Positionen einer Sequenz und eignet sich generell besser z.B. für die Zuordnung einer Funktion eines Proteins, als die sehr kurzen Muster der PROSITE-Datenbank, die aufgrund ihrer Kürze sehr viele Falsch-Positive Treffer erzielen können. Je kürzer ein Muster ist, desto größer ist die Wahrscheinlichkeit, dass mit Hilfe dieses Musters Sequenzen unbekannter Funktion klassifiziert werden

können. Damit steigt auch die Gefahr, Falsch-Positive Treffer zu erzielen. Die Sensitivität eines Sequenzmusters ist nicht nur von der Länge, sondern auch von der Beschaffenheit des Musters abhängig (vgl. Abschnitt 2.8.2).

Die meisten Muster der PROSITE-Datenbank bestehen aus Aminosäuren des aktiven Zentrums [115]. Wie das Beispiel im Abschnitt 3.1 zeigt, liegt das PROSITE-Muster für EC 3.5.99.6 innerhalb der automatisch generierten Muster, da wie erwähnt, die automatisch generierten Muster dieser Arbeit wesentlich länger sind, als Muster aus PROSITE. Für die Enzyme aus Beispiel 3.2 sind in der PROSITE-Datenbank keine Muster vorhanden. Eines von zwei Mustern aus der PROSITE-Datenbank für EC 6.3.5.5 liegt wie im ersten Beispiel innerhalb des generierten Musters.

4.1.7 Beschaffenheit der Muster

Die meisten Muster wurden bei E-Wert 10^{-2} generiert. Da dies die Spitze eines jeden Clusterbaums ist, wurde festgelegt, dass an dieser Stelle stets ein Sequenzalignment erstellt und daraus ein Sequenzmuster extrahiert wird. Dem Diagramm 3-9 ist die Abhängigkeit der erstellten Muster vom E-Wert zu entnehmen. Die meisten Muster wurden bei relativ hohen E-Werten generiert. Die Ursache hierfür ist, dass sich Sequenzen stärker clustern, je größer der E-Wert ist. Mit steigendem E-Wert erhöhen sich sowohl die Wahrscheinlichkeit, dass sich mehr als eine Sequenz im Cluster befindet, als auch auch die Wahrscheinlichkeit, dass aus Sequenzen dieses Clusters ein Muster erstellt wird.

Bei E-Wert 10^{-2} existieren viele Cluster, die nur aus wenigen Sequenzen bestehen (vgl. Diagramm 3-7). Die direkte Folge dieses Umstands ist die Generierung von Mustern aus meist wenigen Sequenzen. Der Großteil, mehr als 70% aller Muster wurde aus bis zu 10 Sequenzen generiert (vgl. Diagramm 3-10). Ist die Anzahl der zur Erstellung eines Musters zu Grunde liegenden Sequenzen gering, nimmt auch die Länge der Muster zu (vgl. Diagramm 3-12). Und da ein Großteil der Muster aus relativ wenigen Sequenzen generiert wurde, sind die erstellten Muster relativ lang. Annähernd 50000 Muster bestehen aus 100 bis 300 Positionen (vgl. Diagramm 3-11).

4.1.8 Qualität der Muster

Anhand von Beispielen im Abschnitt 3 konnte unter anderem die Qualität der Muster eingeschätzt werden. Dazu wurden exemplarisch für den kompletten Datensatz drei Sequenzmuster für verschiedene EC-Nummern aus drei Clusterbäumen ausgewählt und die Musterpositionen in den 3D-Strukturen der jeweiligen Enzyme markiert. Durch farbige Markierungen der Musterpositionen in den dreidimensionalen Moleküldarstellungen der PDB-Datenbank und Vergleiche mit in der Literatur beschriebenen Funktionen wichtiger Aminosäuren, wurde die Aussagekraft der extrahierten Muster bestimmt.

Bei der Untersuchung der Beispiele zeigte sich, dass die Muster meist von guter Qualität sind, d.h. Musterpositionen sind meist Aminosäuren der aktiven oder allosterischen Zentren oder sind Aminosäuren, die dabei beteiligt sind, den Kontakt zwischen Domänen zu gewährleisten. Dies konnte für alle drei Beispiele des Abschnitts 3 gezeigt werden.

Das erste untersuchte Muster für EC 3.5.99.6 wurde aus Sequenzen erstellt, die bei E-Wert 10^{-94} geclustert wurden. Die Aminosäuren des Musters, das aus 131 Positionen besteht, konzentrieren sich stark auf den Bereich des aktiven und des allosterischen Zentrums des Enzyms, wie aus der Darstellung in Abbildung 3-7 hervorgeht. Das zweite Muster für EC 3.5.99.6 wurde aus Sequenzen erstellt, die bei E-Wert 10^{-59} geclustert wurden. Das Sequenzmuster besteht aus 40 Positionen. Anhand der Abbildung 3-8 ist die deutliche Gruppierung der Aminosäuren des Musters um das aktive Zentrum zu erkennen. Mit der Verkürzung dieses Musters im Vergleich mit dem zuvor beschriebenen Muster für dieses Enzym, sind Musterpositionen, die zuvor peripher lagen, in diesem Muster nicht mehr vorhanden. Das automatisch erstellte Muster, das bei E-Wert 10^{-59} erstellt wurde, ähnelt aufgrund der Positionierung der Aminosäuren stark dem manuell erstellten Muster der PROSITE-Datenbank (vgl. Abbildungen 3-6 und 3-8).

Auch im zweiten Beispiel konzentrieren sich die Positionen der Sequenzmuster auf den Bereich des aktiven Zentrums (vgl. Abbildung 3-14). Mit größer werdendem E-Wert, werden die Sequenzmuster kürzer. Je kürzer die Sequenzmuster werden, desto stärker ist die Konzentrierung der Musterpositionen auf den Bereich des aktiven Zentrums (vgl. Abbildungen 3-14 und 3-15). Die Markierungen der automatisch erstellten Sequenzmuster für die Enzyme L-Fuculose-Phosphat Aldolase und L-Ribulose-5-Phosphat 4-Epimerase ähneln stark den Markierungen der Aminosäuren, die während der Katalyse das Zink-Ion und das Substratmolekül binden (vgl. Abbildungen 3-10 und 3-15).

Das untersuchte Muster des dritten Beispiels stammt aus der Sequenz der großen Untereinheit des Enzyms Carbamoylphosphat Synthase. Obwohl die große Untereinheit zwei aktive Zentren enthält, konzentrieren sich die Musterpositionen nicht auf diese Bereiche (vgl. Abbildungen 3-21 und 3-24). Anders als in den Beispielen zuvor, liegen die Aminosäuren im Bereich der großen Untereinheit, die in Kontakt mit der kleinen Untereinheit steht. Teilweise sind im Muster Aminosäuren konserviert, die an der Bildung der intramolekularen Tunnel beteiligt sind.

4.1.9 Untersuchung von Richtig-Positiven und Falsch-Positiven Mustertreffern

In der Datenbank sind 118947 Sequenzmuster vorhanden. Eine Einschätzung der biologischen Bedeutung dieser Muster ist auf verschiedenen Wegen möglich. Auf der einen Seite kann aufgrund von Literaturrecherche bestimmt werden, welche Aminosäuren für die Funktion eines Enzyms von besonderer Bedeutung sind. Dies sind meist Aminosäuren, die direkt bei der katalytischen Reaktion beteiligt, oder Aminosäuren, die besonders wichtig bei der Etablierung einer dreidimensionalen Anordnung der Sequenz sind. Da diese Aminosäuren und die Verteilung dieser Aminosäuren innerhalb der Sequenz typisch für diese Enzyme sind, sollten diese im Muster vorhanden sein, um Sequenzen unbekannter Funktion mit diesem Muster sicher klassifizieren zu können. Diese manuelle Einschätzung der Qualität eines Musters ist für fast 120000 Muster nicht möglich. Stattdessen muss die Qualität der Muster automatisiert bestimmt werden. Einen Hinweis auf die biologische Signifikanz eines Musters liefert die Anzahl Treffer, die ein Muster erzielt, wenn mit diesem Muster gegen eine Sequenzdatenbank gesucht wird. Dieses Verfahren wurde in dieser Arbeit genutzt. Für jedes Muster wurde die Anzahl der erzielten Richtig-Positiven und Falsch-Positiven Treffer in der Datenbank gespeichert. Dabei gilt ein Treffer als Richtig-Positiv, wenn die getroffene Sequenz die gleiche EC-Nummer besitzt, wie das Muster. Besitzt die getroffene Sequenz keine EC-Nummer oder unterscheidet sich die EC-Nummer von der EC-Nummer des Musters, gilt der Treffer als Falsch-Positiv. Eine detailierte Definition und weitere Einzelheiten zur Bestimmung von Richtig-Positiven und Falsch-Positiven Treffern finden sich in Abschnitt 2.9 im Teil Daten, Algorithmen und Methoden.

Wie im bereits in Abschnitt 4.1.7 diskutiert wurde, basieren die meisten Muster auf relativ wenigen Sequenzen. Die daraus resultierenden langen Muster sind sehr spezifisch. Mehr

als 100000 Muster haben mehr oder gleich viele Richtig-Positive Treffer wie Falsch-Positive Treffer (vgl. Tabelle 3-2). 57183 Muster haben keine Falsch-Positiven Treffer. 3075 Muster haben mehr Richtig-Positive Treffer, als aufgrund der Anzahl der Sequenzen, aus denen das Muster erstellt wurde, zu erwarten war und haben keine Falsch-Positiven Treffer. Anhand dieser Ergebnisse kann zusammenfassend festgehalten werden, dass die Muster sehr spezifisch sind, d.h. es gibt wenige Falsch-Positive Treffer, dafür treffen die Muster aber auch meist nur die Sequenzen, die die Grundlage für die Erstellung der Muster bildeten. Aufgrund der Länge der Muster und der damit verbundenen hohen Spezifität muss im Einzelfall geprüft werden, ob Sequenzen unbekannter Funktion mit diesen langen Mustern eindeutig klassifiziert werden können.

4.1.10 Abhängigkeit der Anzahl Falsch-Positiver Treffer von der Musterlänge

Es besteht ein Zusammenhang zwischen der Anzahl Richtig-Positiver und Falsch-Positiver Treffer und der Musterlänge der Muster (vgl. Diagramm 3-16 und Diagramm 3-17). Demnach steigt die Anzahl aller Richtig-Positiver Treffer, je kürzer das Muster ist. Gleichzeitig nimmt auch die Anzahl aller Falsch-Positiven Treffer stark zu, je kürzer das Muster ist. Diese zunächst trivial erscheinenden Aussagen sind bei genauerer Betrachtung von großer Bedeutung, denn durchschnittlich 1200 Falsch-Positive Treffer pro Muster für Muster bis 20 Positionen deuten nicht auf biologisch signifikante Muster hin. 25 Muster, die aus 8 Musterpositionen bestehen, erzielen jeweils über 19000 Falsch-Positive Treffer. Es muss in diesem Zusammenhang erwähnt werden, dass die Häufigkeit von Falsch-Positiven Treffern nicht nur von der Musterlänge, sondern auch von der Beschaffenheit der Muster abhängt. So erzielen Muster, die Positionen für statistisch häufig vorkommende Aminosäuren enthalten, mehr Falsch-Negative Treffer, als Muster, die statistisch seltene Aminosäuren enthalten, obwohl beide Muster gleich lang sind. Und auch flexible Lücken beeinflussen das Ausmaß von Falsch-Positiven Treffern (vgl. Abschnitt 2.8.2). Eine pauschale Definition von „guten“ oder „schlechten“ Mustern ist daher anhand von formalen Angaben wie die Länge der Muster schwierig, da viele Faktoren auf die Anzahl der erzielten Falsch-Positiven Treffer Einfluss haben. Dennoch konnte in der statistischen Auswertung gezeigt werden, dass kürzere Muster in der Regel eine größere Anzahl Falsch-Positiver Treffer erzielen, als längere Muster.

4.1.11 Bestimmung der Musterqualitäten der Beispiele

Anhand von Beispielen im Ergebnisteil wurde untersucht, welche Sequenzen von bestimmten Mustern getroffen wurden, die als Falsch-Positiv deklariert wurden. In der Tabelle 3-17 sind alle Falsch-Positiven Treffer für das Muster für EC 3.5.99.6, Clusterbaum 19905, E-Wert 10^{-59}, zusammengefasst. Das Muster, das aus 40 Musterpositionen besteht, erzielte 82 Richtig-Positive und 34 Falsch-Positive Treffer. Das Muster für das Enzym Glucosamin-6-Phosphat Deaminase traf 25 Sequenzen der Glucosamin-6-Phosphat Isomerase, vier Sequenzen der Glucosamin/Galactosamin-6-Phosphat Isomerase und zwei Sequenzen der N-Acetylglucosamin-6-Phosphat Isomerase. Die verbliebenen drei Falsch-Positiven Treffer verteilen sich auf jeweils eine Sequenz der 6-Phosphogluconolactonase, einem Hypothetischen Protein und einer Putativen Isomerase. Bei allen als Falsch-Positiven bestimmten Sequenzen handelt es sich um Sequenzen für Enzyme. Das nicht näher bezeichnete Hypothetische Protein Q2G0K8 gehörte zum Zeitpunkt der Erstellung des Musters zu der Datenbank TrEMBL. In der aktualisierten Version der Datenbank gehört die Sequenz zu SWISS-PROT und trägt die EC-Nummer 3.5.99.6. Auch die Putative Isomerase trägt nun die EC-Nummer 3.5.99.6. 31 von 34 Enzyme der Falsch-Positiven Treffer nutzen Glucosamin-6-Phosphat als Eduktmolekül, zwei Enzyme nutzen N-Acetylglucosamin, ein Enzym nutzt 6-Phosphogluconolacton als Eduktmolekül. Diese Moleküle sind dem Eduktmolekül des Enzyms 3.5.99.6, Glucosamin-6-Phosphat, zu über 90% ähnlich. Dieser Wert bezieht sich auf die größte gemeinsame Substruktur und wurde mit dem c-MCS Algorithmus bestimmt. Zusammenfassend kann festgehalten werden, dass es sich bei allen Falsch-Positiven Treffern um Sequenzen von Enzymen handelt. Alle Enzyme verwenden identische oder sehr ähnliche Edukte, wie das Enzym Glucosamin-6-Phosphat Deaminase. Auffällig ist auch, dass vorwiegend Isomerasen getroffen werden. Andere Enzyme aus anderen EC-Klassen, die Glucosamin-6-Phosphat als Eduktmolekül nutzen, wie z.B. Glucosamin-Phosphat N-Acetyltransferase, EC 2.3.1.4, werden mit dem Muster für EC 3.5.99.6 nicht getroffen.

Im zweiten Beispiel werden die Falsch-Positiven Treffer für EC 4.1.2.17, L-Fuculose-Phosphat Aldolase, untersucht. Das Muster für dieses Enzym wurde aus 26 Sequenzen generiert, die bei E-Wert 10^{-14} geclustert wurden. Das aus 30 Positionen bestehende Muster erzielte 26 Richtig-Positive und 20 Falsch-Positive Treffer. Als Falsch-Positiv deklarierte Treffer sind Sequenzen der Enzyme L-Fuculose-1-Phosphat Aldolase,

Putative L-Fucolose-Phosphat Aldolase, bzw. die Sequenz für das Enzym L-Fuculose-Phosphat Aldolase. Da das Muster aus Sequenzen des Enzyms L-Fucolose-Phosphat Aldolase entstanden ist, sind diese Treffer nicht als Falsch-Positiv einzustufen, sondern gelten nach manueller Auswertung der Treffer als Richtig-Positiv. Die Treffer wurden als Falsch-Positiv deklariert, da für diese Enzyme in TrEMBL keine EC-Nummer angegeben wurde (vgl. Abschnitt 2.9). Zusätzlich wurde eine Sequenz des Enzyms Ribulose-5-Phosphat 4-Epimerase getroffen. Ein Enzym, dessen Sequenzen in diesem Clusterbaum vorhanden sind. Zwei weitere Falsch-Positive Treffer sind ein Treffer der Sequenz A1BAW5, ein Protein aus der Fis Familie und ein Treffer der Sequenz Q1YJP3, eine nicht näher bezeichnete Aldolase.

Im dritten Beispiel werden die Falsch-Positiven Treffer des Sequenzmusters für das Enzym Carbamoylphosphat Synthase, EC 6.3.5.5 untersucht. Das Muster, das auf Grundlage von 24 Sequenzen bei E-Wert 10^{-84} erstellt wurde, besteht aus 210 Positionen. Bei der Suche gegen SWISS-PROT und TrEMBL wurden 43 Richtig-Positive und 19 Falsch-Positive Treffer erzielt. Alle Falsch-Positiven Treffer dieses Musters stammen aus der Datenbank TrEMBL. Dabei werden durchaus Sequenzen des Enzyms Carbamoylphosphat Synthase getroffen, denn bei 15 dieser Treffer lautet die Bezeichnung "Carbamoyl-phosphate synthase, large subunit", zwei Treffer tragen die Bezeichnung "Carbamoyl-phosphate synthase, large chain", bei einem Treffer lautet die Bezeichnung "PyrAB (Carbamoyl-phosphate synthetase) (Catalytic subunit)". Ein als Falsch-Positiv ermittelter Treffer wird als "CarB" beschrieben, mit, laut TrEMBL, ausgewiesener Sequenzähnlichkeit zur Sequenz der Carbamoylphosphat Synthase. Alle Treffer werden als Falsch-Positiv eingestuft, da in der Beschreibung der Sequenz keine EC-Nummer vorhanden ist (vgl. Abschnitt 2.9).

Besonders durch das zweite und dritte Beispiel wurde deutlich, dass nicht alle als Falsch-Positiv deklarierten Treffer tatsächlich Falsch-Positiv sind. Der Grund dafür sind fehlende EC-Nummern in TrEMBL und fehlende konkrete Angaben in den Beschreibungen der Sequenzen. Aus welchen Gründen bei bestimmten Sequenzen keine EC-Nummern in TrEMBL angegeben wurden, kann nicht eindeutig geklärt werden. Möglicherweise sind die Annotationen fehlerhaft, bzw. die Annotation schien zu unsicher, um bestimmten Sequenzen EC-Nummern und damit eine Funktion zuweisen zu können. Nur anhand der EC-Nummern wurde bestimmt, ob ein Treffer Richtig-Positiv oder Falsch-Positiv ist. In SWISS-PROT sind diese fehlenden Angaben nicht vorstellbar, denn

SWISS-PROT wird im Gegensatz zu TrEMBL gut dokumentiert (vgl. Abschnitte 2.11.1 und 2.11.2). Sollte daher ein Falsch-Positiver Treffer bei der Suche in SWISS-PROT erzielt werden, ist es sehr wahrscheinlich, dass der Treffer tatsächlich Falsch-Positiv ist. Treffer auf Sequenzen mit den Bezeichnungen von z.B. "Hypothetical Protein" oder "Putative Isomerase" sind nicht Falsch-Positiv, sondern aufgrund der unklaren Annotation einfach zu ignorieren. Da diese Angaben keine EC-Nummern enthalten, wurden diese Treffer als Falsch-Positiv gewertet. Solche Sequenzen können über reguläre Ausdrücke identifiziert werden (vgl. Abschnitt 4.2.3).

Die Beispiele verdeutlichen die hohe Qualität der Sequenzmuster. Viele und teilweise alle Sequenzmuster trafen Sequenzen von Enzymen. Viele der als Falsch-Positiv deklarierten Treffer stellten sich nach manueller Begutachtung als Richtig-Positiv heraus. Dies ist nicht auf fehlerhafte Sequenzmuster, sondern auf fehlende Angaben in TrEMBL zurückzuführen. Wird bei der Gesamtbeurteilung der Musterqualitäten die Anzahl der erstellten Muster, die Quote der erzielten Richtig-Positiven Treffer, die Vergleiche der Muster mit den Mustern der PROSITE-Datenbank und die Positionierung der Aminosäuren der Muster in den dreidimensionalen Strukturdarstellungen berücksichtigt, stellt der in dieser Arbeit vorgestellte Ansatz zur Identifizierung, automatischen Generierung und Analyse von konservierten Sequenzmustern eine sinnvolle Alternative zur Datenbank PROSITE dar. Insbesondere vor dem Hintergrund, da die Sequenzmuster in diesem Ansatz automatisch generiert werden und die Datenbank leicht aktualisiert werden kann. Die Datenbank kann damit schnell und verlässlich Daten in guter Qualität einem breiten Publikum zur Verfügung stellen.

4.1.12 Diskussion der Ergebnisse der Untersuchung der größten gemeinsamen Teilstruktur von Edukten/Produkten/Co-Substrate

Mit Hilfe des c-MCS Algorithmus‘ wird die größte gemeinsame Teilstruktur der Moleküle untersucht, die bei den verglichenen Reaktionen beteiligt sind. Zudem wird das Ergebnis der Clusterung, das auf Sequenzähnlichkeit basiert, mit der Clusterung verglichen, die auf identischen R-Matrizen basiert. Beide Verfahren dienen dazu, eine Antwort auf die Frage zu geben, inwiefern es möglich ist, von einer Sequenzähnlichkeit auf eine gleiche Funktion schließen zu können. Als Unterscheidungsmerkmal von Enzymen wird das EC-System genutzt, das Enzyme aufgrund ihrer katalysierten Reaktion und nicht aufgrund von Sequenzähnlichkeit klassifiziert.

In der Clusteranalyse wurden homologe Enzymsequenzen untersucht. Aufgrund dieser Tatsache ist es sehr wahrscheinlich, dass Sequenzen, die aufgrund von Sequenzähnlichkeit geclustert wurden, eine ähnliche Struktur [117, 118] und eine ähnliche Funktion [52] aufweisen, falls die Sequenzidentitäten über den Bereich der *twilight zone* liegen. Die *twilight zone* stellt den Bereich dar, in dem Sequenzidentitäten unter 40% liegen. Untersuchungen von Wilson *et al.* [119], Devos und Valencia [120] und Todd *et al.* [121] zeigten die Möglichkeit auf, von Sequenzähnlichkeit auf ähnliche Funktion zu schließen, falls eine signifikante Sequenzähnlichkeit vorliegt.

Bei verschiedenen E-Werten wurden paarweise EC-Kombinationen untersucht, in welchem Ausmaß die Edukte, Produkte bzw. Co-Substrate dieser Reaktionen übereinstimmen. Zu diesem Zweck wurde der c-MCS Algorithmus von Markus Leber [86] verwendet. Bei E-Wert 10^{-181} wurden 42 EC-Kombinationen verglichen. Mit ansteigendem E-Wert steigt die Anzahl der paarweise untersuchten EC-Kombinationen an. Bei E-Wert 10^{-2} konnten 2462 EC-Kombinationen verglichen werden (vgl. Abschnitt 2.10.5 und Tabelle 3-7).

Wie man dem Diagramm 3-19 entnehmen kann, sind bei jedem E-Wert die bei den katalysierten Reaktionen beteiligten Moleküle der untersuchten EC-Kombinationen zum größten Teil identisch. Bei E-Wert 10^{-181} ist bei 78,6% aller verglichenen Reaktionen mindestens ein Molekül bei beiden Reaktionen identisch. Bis E-Wert 10^{-120} steigt der Wert leicht auf 82,2% an, danach fällt der Wert bis E-Wert 10^{-2} auf 65% ab. Mit steigendem E-Wert nimmt der Anteil identischer Moleküle der verglichenen Reaktionen, bis auf den leichten Anstieg bei E-Wert 10^{-120}, deutlich ab.

Das Diagramm 3-20 stellt die Abhängigkeit der prozentualen größten gemeinsamen Teilstruktur der bei den untersuchten Reaktionen enthaltenen Moleküle vom E-Wert dar. Es ist zu erkennen, dass falls eine Übereinstimmung detektiert wurde, diese meist bei über 40% liegt. Geringere prozentuale Übereinstimmungen kommen bei allen E-Werten nur selten vor. Tendenziell ist der Anteil der geringeren prozentualen Übereinstimmung bei hohen E-Werten höher, als bei niedrigen E-Werten. Je niedriger der E-Wert ist, desto höher ist der prozentuale Anteil der größeren gemeinsamen Teilstruktur.

4.1.13 Diskussion der Einordnung der untersuchten EC-Kombinationen in die Clusterung nach gleichen R-Strings

Markus Leber hat in seiner Dissertation verschiedene EC-Subsubklassen nach gleichen R-Strings geclustert (vgl. Abschnitt 2.10). Dieser auf diese Weise erhaltene Datensatz und die damit erhaltenen EC-Kombinationen wurden mit der Clusterung von homologen Sequenzen dieser Arbeit verglichen. Es wurde untersucht, in welchem Maße EC-Kombinationen, die aufgrund von Sequenzähnlichkeit geclustert wurden, mit EC-Kombinationen, die aufgrund identischer R-Matrizen geclustert wurden, übereinstimmen. Die Ergebnisse dieser Untersuchung werden in Tabelle 3-8 und Diagramm 3-21 dargestellt. Bei verschiedenen E-Werten, von 10^{-181} bis 10^{-2}, wurden insgesamt 4282 EC-Kombinationen auf Zugehörigkeit der Reaktionen zu Clustern identischer R-Strings untersucht. Bei E-Wert 10^{-181} gehörten 5 von 45 EC-Kombinationen dem gleichen R-Cluster an. Mit steigendem E-Wert steigt die Anzahl der untersuchten EC-Kombinationen an. Es ist zu sehen, dass mit steigendem E-Wert der prozentuale Anteil der EC-Kombinationen, die zum gleichen R-Cluster gehören, abnimmt (vgl. Diagramm 3-21). Nur zwischen den E-Werten 10^{-181} bis 10^{-160} steigt der Anteil leicht an. Dies könnte an der geringen Anzahl der untersuchten Kombinationen und dem damit verbundener Abweichung bei E-Wert 10^{-181} liegen.

Als Schlussfolgerung der Untersuchungen der größten gemeinsamen Teilstrukturen von Molekülen und der Vergleich der Clusterung der Enzyme aufgrund identischer R-Strings kann festgehalten werden, dass die in dieser Arbeit erzielten Ergebnisse einen Zusammenhang zwischen Sequenz- und Funktionsähnlichkeit bestätigen.

4.1.14 Darstellung der Clusterbäume

Die Baumstrukturen aller Clusterbäume wurden in den Tabellen *tree*, *edges* und *ecnumbers* gespeichert. Die Informationen dieser Tabellen können mit Hilfe des Programms yEd aus dem Programmpaket yFiles grafisch dargestellt werden (vgl. Abschnitt 2.13.3).

Da Clusterbäume aus mehreren tausend Knoten bestehen können, ist die graphische Darstellung von Clusterbäumen sehr wichtig. Die Knoten enthalten Informationen über E-Wert, EC-Nummern, enthaltene Sequenzen, Clusternummer, Baumnummer und ggf. Muster. Mittels der grafischen Darstellung der Clusterbäume ist es möglich, Cluster, aus

deren Sequenzen Muster erstellt wurden, schnell zu erkennen. Eine Analyse der Beziehungen der Cluster untereinander ist auf diese Weise möglich.

Die selbstgeschriebenen Programme, die im Rahmen dieser Arbeit optimiert wurden, erstellen aus Informationen der Datenbank eine Datei im gml-Format (vgl. DVD). Dieses Dateiformat ist ein Standardformat zur Speicherung graphischer Informationen und kann von verschiedenen Programmen, z.B. von yEd, die zur graphischen Darstellung fähig sind, geladen werden.

Es gibt zwei Darstellungsarten der Clusterbäume. Diese unterscheiden sich in der Anordnung der Clusterknoten und nicht im Inhalt der Knoten. Die gewünschten Darstellungen sind vom Benutzer frei wählbar und haben folgende Vor- und Nachteile:

1. Die optimierte hierarchische Darstellung Baumdarstellung nach yEd:

- Vorteil: Der Baum wird übersichtlicher dargestellt.
- Nachteil: yEd schließt bei der Anordnung der Clusterknoten vorhandene E-Werte nicht mit ein. Zwar befinden sich niemals Clusterknoten mit niedrigen E-Werten über Clusterknoten mit hohen E-Werten. Dennoch können unterschiedliche E-Werte nebeneinander liegen. Eine Abhilfe liefert eine farbige Markierung unterschiedlicher E-Werte.

2. Cluster in hierarchischer Abbildung nach dem E-Wert:

- Vorteil: Man erhält einen schnellen Überblick über die Stellen des Clusterbaums, an denen sich Knoten in Abhängigkeit des E-Wertes befinden.
- Nachteil: Clusterbäume, die nur aus wenigen Knoten bestehen und sich an unterschiedlichen E-Werten befinden, sind sehr lang und unübersichtlich.

Ein genereller Nachteil bei der Darstellung der Clusterbäume mit Hilfe von yED ist, dass Text in den Knoten nur einfarbig dargestellt werden kann. Es wurde überlegt, Positionen im Muster farbig zu markieren. Da yEd zur mehrfarbigen Darstellung der Muster nicht fähig ist, wurde dieser Schritt nicht realisiert. Stattdessen wurden Muster mit farbig hervorgehobenen Aminosäuren manuell erstellt und im Abschnitt 3 dargestellt.

4.2 Ausblick

4.2.1 Verbesserung der Alignments

CLUSTAL W hat sich für die Durchführung von Alignments bewährt, es konnte aber gezeigt werden, dass mit steigender Sequenzzahl die Identifizierung von konservierten Positionen schwierig ist. Eine Verbesserung der Alignments könnte durch individuelle Einstellungen in CLUSTAL W erreicht werden. Dazu müssen genaue Informationen über die zu untersuchenden Sequenzen vorhanden sein, die unter Umständen aus den Beschreibungen aus SWISS-PROT und TrEMBL gewonnen werden können. Mit Hilfe von Angaben über Sequenzlänge oder erwartete Ähnlichkeit, kann die Qualität der Alignments verbessert werden. Für diesen Ansatz muss ein Verfahren entwickelt werden, das Angaben automatisch auslesen und an CLUSTAL W übergeben kann. Dies könnte über die Verwendung von individuellen regulären Ausdrücken realisiert werden.

4.2.2 Erhöhung der Mindestanzahl von Sequenzmustern

Für zukünftige Versionen dieser Datenbank wäre es sinnvoll, eine Mindestsequenzanzahl für Muster zu bestimmen, denn je weniger Sequenzen einem Muster zu Grunde liegen, desto länger wird das Muster. Es ist fraglich, ob tatsächlich alle Aminosäuren biologisch relevant sind, wenn ein Muster 90% aller Aminosäuren der Sequenzen repräsentiert. Mit zunehmenden Sequenzzahlen im Cluster werden die erstellten Muster deutlich kürzer. Nach den vorliegenden Ergebnissen dieser Arbeit wäre eine Sequenzanzahl von mindestens 20 Sequenzen ratsam, obwohl durch diese Maßnahme die Anzahl der erstellten Muster deutlich abnehmen wird, da die meisten Cluster selbst bei E-Wert 10^{-2} relativ klein sind (vgl. Diagramm 3-7). Da aber zukünftige Sequenzdatenbanken an Größe zunehmen werden, nimmt auch die Wahrscheinlichkeit zu, dass durch Clusterung von einer größeren Anzahl Sequenzen tendenziell größere Cluster entstehen werden.

4.2.3 Änderung der Identifizierung von Richtig-Positiven und Falsch-Positiven Treffern

In der vorliegenden Arbeit wurde die Bewertung von Sequenztreffern von Sequenzmustern ausschließlich anhand des EC-Systems vorgenommen (vgl. Abschnitt 2.9). Dieser Ansatz ist nur dann erfolgsversprechend, falls in den untersuchten Datenbanken verlässliche

Informationen über EC-Nummern zur Verfügung gestellt werden. Wie die untersuchten Beispiele zeigen, wurden viele Treffer als Falsch-Positiv bewertet, weil EC-Nummern in den Beschreibungen der Sequenzen, nicht vorhanden sind. Diese Treffer hätten als Richtig-Positiv eingeschätzt werden müssen, da Sequenzen von Enzymen aus der gleichen Familie stammen, wie die Sequenzen des Musters. Als Konsequenz der fehlenden Informationen in TrEMBL ist es in Zukunft vorstellbar, mit regulären Ausdrücken nach Namen des Enzyms zu suchen und die Beurteilung von Richtig-Positiven oder Falsch-Positiven Treffer anhand von Enzymnamen durchzuführen und nicht ausschließlich anhand von EC-Nummern. Das EC-System kann bei der Beurteilung von Treffern behilflich sein.

4.2.4 Reduzierung ähnlicher Muster

Um die Menge der erhaltenen Sequenzmuster zu reduzieren, ist es vorstellbar, die Sequenzmuster eines Clusterbaums, die eine höhere Anzahl Falsch-Positiver Treffer liefern, als vergleichbare Muster des Clusterbaums, zu löschen. Die EC-Zusammensetzung und die Anzahl Richtig-Positiver Treffer der Muster müssen gleich sein. Durch die Löschung von Mustern aus dem gleichen Clusterbaum wird gewährleistet, dass Sequenzmuster gleiche Sequenzbereiche abdecken.

4.2.5 Anhebung der Mindestlänge von Mustern

Obwohl die durchschnittlich größte Anzahl Richtig-Positiver Treffer mit Mustern erzielt wurden, die 8 bis 20 Musterpositionen lang sind, ist es doch für zukünftige Versionen der Datenbank sinnvoll, die Mindestlänge von 8 Musterpositionen auf 15 bis 20 Positionen anzuheben, da ein Zusammenhang zwischen der Anzahl Falsch-Positiver Treffer und der Musterlänge besteht (vgl. Diagramm 3-17). Eine Alternative dazu wäre, dass die Mindestlänge eines Musters bei 8 Positionen belassen wird, aber die Muster, die besonders viele Falsch-Positive Treffer erzielen, aus der Datenbank gelöscht werden.

4.2.6 Identifizierung eines Musters auf einer bestimmten Sequenz

Ein Sequenzmuster gibt Auskunft über konservierte Positionen einer Proteinsequenz. Da oft die Sequenz eines bestimmten Organismus‘ untersucht wird, ist es erforderlich, die Positionen des Musters auf der Enzymsequenz dieses Organismus zu identifizieren, z.B.

für Mutationsexperimente oder die Markierung der Positionen in einer PDB-Datei. Die Zuordnung geschieht manuell und wird umso schwieriger, je länger das Muster oder die Enzymsequenz ist. Ein Programm, das Sequenzpositionen des Musters automatisiert identifiziert und markiert, wäre sehr sinnvoll.

4.2.7 Entwicklung einer Applikation

Um die Daten einem großen Publikum zugänglich zu machen, könnten die Datenbanken und Programme im Internet veröffentlicht werden. Vorstellbar ist zum Beispiel eine Webapplikation, die es ermöglicht, nach EC-Nummern oder Uniprot Accession-Nummern zu suchen und Clusterbäume mit den entwickelten Programmen selbst zu erstellen. Dies wäre eine sinnvolle Ergänzung zu bestehenden Datenbanken wie BRENDA oder PROSITE.

4.2.8 Optimierung der Musterdarstellung

Die Darstellung der Clusterbäume mit yEd aus der Datenbank konnte automatisiert werden. Die Clusterknoten enthalten alle wichtigen Informationen. Eine farbige Markierung von Musterpositionen aus dem Splitcluster, die auch im Muttercluster existieren, würde den Vergleich der Muster aus Mutter- und Splitcluster besonders bei langen Sequenzen extrem vereinfachen. Eine farbige Markierung der Positionen ist grundsätzlich möglich, kann aber von yEd nicht dargestellt werden. Eine Verbesserung der Darstellung würde vereinfacht werden, es ist aber kein Programm vorhanden, das den Ansprüchen genügt. Möglicherweise kann ein selbstentwickeltes Programm dieses Problem lösen.

Literaturverzeichnis

[1] Bairoch, A. (1993):
The PROSITE dictionary of sites and patterns in proteins, its current status.
Nucleic Acids Res., 21(13):3097-3103.

[2] Bork, P. & Koonin, E.V. (1996):
Protein sequence motifs.
Curr Opin Struct Biol., 6(3):366-376. Review.

[3] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990):
Basic local alignment search tool.
J Mol Biol., 215(3):403-410.

[4] Pearson, W.R. & Lipman, D.J. (1988):
Improved tools for biological sequence comparison.
Proc Natl Acad Sci U S A. 85(8):2444-2448.

[5] Smith, T.F. & Waterman, M.S. (1981):
Identification of common molecular subsequences.
J Mol Biol., 147(1):195-197.

[6] Carillo, H. & Lipman, D. (1988):
The multiple sequence alignment problem in biology.
SIAM J. Appl. Math., 48(5):1073-1082.

[7] Hirosawa, M., Totoki, Y., Hoshida, M. & Ishikawa, M. (1995):
Comprehensive study on iterative algorithms of multiple sequence alignment.
Comput Appl Biosci., 11(1):13-18.

[8] Waterman, M.S., Arratia, R., & Galas, D.J. (1984):
Pattern recognition in several sequences: consensus and alignment.
Bull Math Biol., 46(4):515-527.

[9] Feng, D.F. & Doolittle, R.F. (1987):
Progressive sequence alignment as a prerequisite to correct phylogenetic trees.
J Mol Evol., 25(4):351-360.

[10] Feng, D.F. & Doolittle, R.F. (1996):
Progressive alignment of amino acid sequences and construction of phylogenetic trees from them.
Methods Enzymol., 266:368-382.

[11] Martinez, H.M. (1988):
A flexible multiple sequence alignment program.
Nucleic Acids Res., 16(5):1683-1691.

[12] Wu, T.D. & Brutlag, D.L. (1995):
Identification of protein motifs using conserved amino acid properties and partitioning techniques.
Proc Int Conf Intell Syst Mol Biol., 3:402-410.

[13] Sobel, E. & Martinez, H.M. (1986):
A multiple sequence alignment program.
Nucleic Acids Res., 14(1):363-374.

[14] Edgar, R.C. (2004):
MUSCLE: multiple sequence alignment with high accuracy and high throughput.
Nucleic Acids Res., 32(5):1792-1797.

[15] Edger, R.C. (2004):
MUSCLE: a multiple sequence alignment method with reduced time and space complexity.
BMC Bioinformatics, 5:113.

[16] Schwartz, A.S. & Pachter, L. (2007):
Multiple alignment by sequence annealing.
Bioinformatics, 23(2):e24-e29.

[17] Bray, N. & Pachter, L. (2004):
MAVID: constrained ancestral alignment of multiple sequences.
Genome Res., 14(4):693-699.

[18] Higgins, D.G., Thompson, J.D. & Gibson, T.J. (1996):
Using CLUSTAL for multiple sequence alignments.
Methods Enzymol., 266:383-402.

[19] Notredame, C., Higgins, D.G. & Heringa, J. (2000):
T-Coffee: A novel method for fast and accurate multiple sequence alignment.
J Mol Biol., 302(1):205-217.

[20] Internetquelle (02/2008):
T-Coffee zur interaktiven Eingabe im Internet. Online im Internet.
Adresse: http://www.tcoffee.org.

[21] Attwood, T.K., Beck, M.E., Bleasby, A.J. & Parry-Smith, D.J. (1994):
PRINTS-a database of protein motif fingerprints.
Nucleic Acids Res., 22(17):3590-3596.

[22] Jonassen, I., Collins, J.F. & Higgins, D.G. (1995):
Finding flexible patterns in unaligned protein sequences.
Protein Sci., 4(8):1587-1595.

[23] Jonassen, I. (1997):
Efficient discovery of conserved patterns using a pattern graph.
Comput Appl Biosci., 13(5):509-522.

[24] Internetquelle (02/2008):
Pratt - Pattern Matching. Online im Internet.
Adresse: http://www.ebi.ac.uk/pratt.

[25] Rigoutsos, I. & Floratos, A. (1998):
Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm.
Bioinformatics, 14(1):55-67.

[26] Rigoutsos, I. & Floratos, A. (1998):
Motif Discovery Without Alignment Or Enumeration.
In Proceedings Second Annual ACM International Conference on Computational Molecular Biology (RECOMB 98), New York, NY.

[27] Benson, G. & Waterman, M.S. (1994):
A method for fast database search for all k-nucleotide repeats.
Nucleic Acids Res., 22(22):4828-4836.

[28] Sagot, M.F., Viari, A., Pothier, J. & Soldano, H. (1995):
Finding flexible patterns in a text: an application to three-dimensional molecular matching.
Comput Appl Biosci., 11(1):59-70.

[29] Escalier, V., Pothier, J., Soldano, H. & Viari, A. (1998):
Pairwise and multiple identification of three-dimensional common substructures in proteins.
J Comput Biol., 5(1):41-56.

[30] Roytberg, M.A. (1992):
A search for common patterns in many sequences.
Comput Appl Biosci., 8(1):57-64.

[31] Martinez, H.M. (1983):
An efficient method for finding repeats in molecular sequences.
Nucleic Acids Res., 11(13):4629-4634.

[32] Neuwald, A.F. & Green, P. (1994):
Detecting patterns in protein sequences.
J Mol Biol., 239(5):698-712.

[33] Wang, J.T., Marr, T.G., Shasha, D., Shapiro, B.A. & Chirn, G.W. (1994):
Discovering active motifs in sets of related protein sequences and using them for classification.
Nucleic Acids Res., 22(14):2769-2775.

[34] Wang, L. & Jiang, T. (1994):
On the complexity of multiple sequence alignment.
J Comput Biol., 1(4):337-348.

[35] Kunik, V., Solan, Z., Edelman, S., Ruppin, E. & Horn, D. (2005):
Motif extraction and protein classification.
Proc IEEE Comput Syst Bioinform Conf., 80-85.

[36] Smith, H.O., Annau, T.M. & Chandrasegaran, S. (1990):
Finding sequence motifs in groups of functionally related proteins.
Proc Natl Acad Sci U S A, 87(2):826-830.

[37] Smith, R.F. & Smith, T.F. (1990):
Automatic generation of primary sequence patterns from sets of related protein sequences.
Proc Natl Acad Sci U S A, 87(1):118-122.

[38] Suyama, M., Nishioka, T. & Oda, J. (1995):
Searching for common sequence patterns among distantly related proteins.
Protein Eng., 8(11):1075-1080.

[39] Gonnet, P. & Lisacek, F. (2002):
Probabilistic alignment of motifs with sequences.
Bioinformatics, 18(8):1091-1101.

[40] Tao, T., Zhai, C.X., Lu, X. & Fang, H. (2004):
A study of statistical methods for function prediction of protein motifs.
Appl Bioinformatics, 3(2-3):115-124.

[41] Henikoff, S. & Henikoff, J.G. (1991):
Automated assembly of protein blocks for database searching.
Nucleic Acids Res., 19(23):6565-6572.

[42] Henikoff, S. & Henikoff, J.G. (1994):
Protein family classification based on searching a database of blocks.
Genomics, 19(1):97-107.

[43] Bucher, P. & Bairoch, A. (1994):
A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation.
Proc Int Conf Intell Syst Mol Biol., 2:53-61.

[44] Gribskov, M., McLachlan, A.D. & Eisenberg, D. (1987):
Profile analysis: detection of distantly related proteins.
Proc Natl Acad Sci U S A. 84(13):4355-4358.

[45] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D. & Sonnhammer, E.L. (1999):
Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.
Nucleic Acids Res., 27(1):260-262.

[46] Krogh, A., Brown, M., Mian, I.S., Sjölander, K. & Haussler, D. (1994):
Hidden Markov models in computational biology. Applications to protein modeling.
J Mol Biol., 235(5):1501-1531.

[47] Hart, R.K., Royyuru, A.K., Stolovitzky, G. & Califano, A. (2000):
Systematic and fully automated identification of protein sequence patterns.
J Comput Biol., 7(3-4):585-600.

[48] Henikoff, S., Henikoff, J.G. & Pietrokovski, S. (1999):
Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations.
Bioinformatics, 15(6):471-479.

[49] Brazma, A., Vilo, J., Ukkonen, E. & Valtonen, K. (1997):
Data mining for regulatory elements in yeast genome.
Proc Int Conf Intell Syst Mol Biol., 5:65-74.

[50] Bystroff, C. & Baker, D. (1998):
Prediction of local structure in proteins using a library of sequence-structure motifs.
J Mol Biol., 281(3):565-577.

[51] Bystroff, C. & Shao, Y. (2002):
Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA.
Bioinformatics, 18 Suppl 1:S54-61.

[52] von Grotthuss, M., Plewczynski, D., Ginalski, K., Rychlewski, L. & Shakhnovich, E.I. (2006):
PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics.
BMC Bioinformatics, 7:53.

[53] Bystroff, C., Simons, K.T., Han, K.F. & Baker, D. (1996):
Local sequence-structure correlations in proteins.
Curr Opin Biotechnol., 7(4):417-421. Review.

[54] Lin, K.Y., Wright, J. & Lim, C. (2000):
Conformational analysis of long spacers in PROSITE patterns.
J Mol Biol., 299(2):537-548.

[55] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M., Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J., Silventoinen, V., Studholme, D.J., Vaughan, R. & Wu, C.H. (2005):
InterPro, progress and status in 2005.
Nucleic Acids Res., 33(Database issue):D201-D205.

[56] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. & Sigrist, C.J. (2006):
The PROSITE database.
Nucleic Acids Res., 34(Database issue):D227-D230.

[57] Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygouri, C. (2003):
PRINTS and its automatic supplement, prePRINTS.
Nucleic Acids Res., 31(1):400-402.

[58] Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. & Kahn, D. (2005):
The ProDom database of protein domain families: more emphasis on 3D.
Nucleic Acids Res., 33(Database issue):D212-D215.

[59] Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L. & Bateman, A. (2006):
Pfam: clans, web tools and services.
Nucleic Acids Res., 34(Database issue):D247-D251.

[60] Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006):
SMART 5: domains in the context of genomes and networks.
Nucleic Acids Res., 34(Database issue):D257-D260.

[61] Haft, D.H., Selengut, J.D. & White, O. (2003):
The TIGRFAMs database of protein families.
Nucleic Acids Res., 31(1):371-373.

[62] Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.S., Natale, D.A., Vinayaka, C.R., Hu, Z.Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R.S., Suzek, B.E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J.L., Chung, S., Castro-Alvear, J., Dinkov, G. & Barker, W.C. (2004):
PIRSF: family classification system at the Protein Information Resource.
Nucleic Acids Res., 32(Database issue):D112-D114.

[63] Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001):
Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.
J Mol Biol., 313(4):903-919.

[64] Yeats, C., Maibaum, M., Marsden, R., Dibley, M., Lee, D., Addou, S. & Orengo, C.A. (2006):
Gene3D: modelling protein structure, function and evolution.
Nucleic Acids Res., 34(Database issue):D281-D284.

[65] Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H. & Thomas, P.D. (2005):
The PANTHER database of protein families, subfamilies, functions and pathways.
Nucleic Acids Res., 33(Database issue):D284-D288.

[66] Internetquelle (02/2008):
Die Datenbank Interpro. Online im Internet.
Adresse: ftp://ftp.ebi.ac.uk/pub/databases/interpro.

[67] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S. & Sigrist, C.J. (2007):
The 20 years of PROSITE.
Nucleic Acids Res., 36(Database issue):D245-D249.

[68] Bairoch, A. (1991):
PROSITE: a dictionary of sites and patterns in proteins.
Nucleic Acids Res., 19 Suppl:2241-2245.

[69] Coin, L., Bateman, A. & Durbin, R. (2004):
Enhanced protein domain discovery using taxonomy.
BMC Bioinformatics, 5:56.

[70] Portugaly, E., Linial, N. & Linial, M. (2007):
EVEREST: a collection of evolutionary conserved protein domains.
Nucleic Acids Res., 35(Database issue):D241-D246.

[71] Bradley, P., Kim, P.S. & Berger, B. (2002):
TRILOGY: Discovery of sequence-structure patterns across diverse proteins.
Proc Natl Acad Sci U S A. 99(13):8500-8505.

[72] Brenner, S.E., Chothia, C. & Hubbard, T.J. (1998):
Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.
Proc Natl Acad Sci U S A. 95(11):6073-6078.

[73] Saigo, H., Vert, J.P., Ueda, N. & Akutsu, T. (2004):
Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682-1689.

[74] Chothia, C. & Lesk, A.M. (1986):
The relation between the divergence of sequence and structure in proteins.
EMBO J., 5(4):823-826.

[75] Jain, A., Murty, M. & Flynn, P. (1999):
Data clustering: a review.
ACM Computing Reviews, vol 31, no 3, 264-323.

[76] Piatetsky-Shapiro, G. & Frawley, W.J. (1991):
Knowledge Discovery in Databases.
AAAI/MIT Press 1991.

[77] Krause, A., Stoye, J. & Vingron, M. (2005):
Large scale hierarchical clustering of protein sequences.
BMC Bioinformatics, 6:15.

[78] Yona, G., Linial, N. & Linial, M. (2000):
ProtoMap: automatic classification of protein sequences and hierarchy of protein families.
Nucleic Acids Res., 28(1):49-55.

[79] Enright, A.J. & Ouzounis, C.A. (2000):
GeneRAGE: a robust algorithm for sequence clustering and domain detection.
Bioinformatics, 16(5):451-457.

[80] Aus dem Spring, C. (2006*):*
Identifizierung ähnlicher Reaktionsmechanismen in homologen Enzymen unterschiedlicher Funktion unter Verwendung konservierter Sequenzdomänen.
Dissertation, Universität zu Köln.

[81] Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J.M. & Orengo, C.A. (2007):
The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution.
Nucleic Acids Res., 35(Database issue):D291-D297.

[82] Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C. & Murzin, A.G. (2004):
SCOP database in 2004: refinements integrate structure and sequence family data.
Nucleic Acids Res., 32(Database issue):D226-D229.

[83] 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.

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

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

[86] Leber, M. (2008):
Kodierung enzymatischer Reaktionen.
Dissertation, Universität zu Köln.

[87] Bron, C. & Kerbosch, J. (1973):
Algorithm 457 - Finding all cliques of an undirected graph.
Communications of the ACM 16, 575-577.

[88] Huang, X., Lai, J. & Jennings, S.F. (2006):
Maximum common subgraph: some upper bound and lower bound results.
BMC Bioinformatics. 7 Suppl 4:S6.

[89] Kaufman, L. & Rousseeuw, P.J. (1990):
Finding Groups in Data: an Introduction to Cluster Analysis.
Wiley, New York.

[90] Hartigan, J.A. (1975):
Clustering Algorithms.
Wiley, New York.

[91] Matsuda, H., Ishihara, T. & Hashimoto, A. (1999):
Classifying molecular sequences using a linkage graph with their pairwise similarities,
Theor. Comput. Sci., 210:305–325.

[92] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N. & Suzek, B. (2006):
The Universal Protein Resource (UniProt): an expanding universe of protein information.
Nucleic Acids Res., 34(Database issue):D187-D191.

[93] Rice, P., Longden, I. & Bleasby, A. (2000):
EMBOSS: The European Molecular Biology Open Software Suite.
Trends in Genetics, 16, (6) pp276-277.

[94] Wootton, J.C. (1994):
Non-globular domains in protein sequences: automated segmentation using complexity measures.
Comput Chem., 18(3):269-285.

[95] Wootton, J.C. & Federhen, S. (1996):
Analysis of compositionally biased regions in sequence databases.
Methods Enzymol., 266:554-571.

[96] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997):
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
Nucleic Acids Res., 25(17):3389-3402. Review.

[97] Henikoff, S. & Henikoff, J.G. (1992):
Amino acid substitution matrices from protein blocks.
Proc Natl Acad Sci USA, 89(22):10915-10919.

[98] Karlin, S. & Altschul, S.F. (1990):
Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.
Proc Natl Acad Sci USA, 87(6):2264-2268.

[99] Altschul, S.F. (1991):
Amino acid substitution matrices from an information theoretic perspective.
J Mol Biol., 219(3):555-565

[100] Internetquelle (04/2008):
BLAST Tutorials. Online im Internet.
Adresse: http://www.swbic.org/origin/proc_man/Blast/BLAST_tutorial.html.

[101] Thompson, J.D., Higgins, D.G., & Gibson, T.J. (1994):
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.
Nucleic Acids Res., 22(22):4673-4680.

[102] Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994).
Improved Sensitivity of profile searches through the use of sequence weights and gap excision.
Cabios, 10(1), 19-29.

[103] Internetquelle (04/2008):
Pythonkurs des Forschungszentrums Jülich. Online im Internet.
Adresse: http://www.fz-juelich.de/jsc/neues/termine/python.

[104] Internetquelle (04/2008):
MySQL: Die populärste Open-Source-Datenbank der Welt. Online im Internet.
Adresse: http://www.mysql.de.

[105] Internetquelle (04/2008):
Das Rechenzentrum der Universität zu Köln. Online im Internet.
Adresse: http://www.uni-koeln.de/rrzk.

[106] Internetquelle (04/2008):
Die Technische Universität Braunschweig. Online im Internet.
Adresse: http://www.tu-braunschweig.de/flw/forschung/schwerpunkte.

[107] Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. & Murzin, A.G. (2008):
Data growth and its impact on the SCOP database: new developments.
Nucleic Acids Res, 36(Database issue):D419-D425.

[108] Internetquelle (04/2008):
Bayerische Akademie der Wissenschaften, Unterlagen zur Erklärung von Regulären Ausdrücken. Online im Internet.
Adresse: http://www.lrz-muenchen.de/services/schulung/unterlagen/regul.

[109] Internetquelle (04/2008):
Das Modul re in python. Online im Internet.
Adresse: http://docs.python.org/lib/module-re.html.

[110] Thoden, J.B., Raushel, F.M., Benning, M.M., Rayment, I. & Holden, H.M. (1999):
The structure of carbamoyl phosphate synthetase determined to 2.1 A resolution.
Acta Crystallogr D Biol Crystallogr., 55(Pt 1):8-24.

[111] Braxton, B.L., Mullins, L.S., Raushel, F.M. & Reinhart, G.D. (1999):
Allosteric dominance in carbamoyl phosphate synthetase.
Biochemistry, 38(5):1394-1401.

[112] Miles, B.W. & Raushel, F.M. (2000):
Synchronization of the three reaction centers within carbamoyl phosphate synthetase.
Biochemistry, 39(17):5051-5056.

[113] Arikawa, S., Kuhara, S., Miyano, S., Shinohara, A. & Shinohara, T. (1991):
A Learning algorithm for elementary formal systems and its experiments on Identification of transmembrane domains.
Proc. 25th Int.Conf. Hawaii International Conference on Information Systems, 675-684.

[114] Sagot, M.F., Viari, A. & Soldano, H. (1995):
A distance-based block searching algorithm.
Proc Int Conf Intell Syst Mol Biol., 3:322-331.

[115] Kasuya, A. & Thornton, J.M. (1999):
Three-dimensional structure analysis of PROSITE patterns.
J Mol Biol., 286(5):1673-1691.

[116] Sonnhammer, E.L., Eddy, S.R. & Durbin, R. (1997):
Pfam: a comprehensive database of protein domain families based on seed alignments.
Proteins, 28(3):405-420.

[117] Rost, B. (1999):
Twilight zone of protein sequence alignments.
Protein Eng., 12(2):85-94.

[118] Chung, S.Y. & Subbiah, S. (1996):
A structural explanation for the twilight zone of protein sequence homology.
Structure, 4(10):1123-1127. Review.

[119] Wilson, C.A., Kreychman, J., & Gerstein, M. (2000):
Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.
J Mol Biol., 297(1):233-249.

[120] Devos, D. & Valencia, A. (2000):
Practical limits of function prediction.
Proteins, 41(1):98-107.

[121] Todd, A.E., Orengo, C.A. & Thornton, J.M.(2002):
Sequence and structural differences between enzyme and nonenzyme homologs.
Structure, (10):1435-1451.

[122] Internetquelle (04/2008):
Gentechnik pro&contra. Online im Internet.
Adresse: http://www.gentech.sbg.ac.at/anwendungen.htm.

[123] Internetquelle (04/2008):
Die europäische Kommission: Verbesserung der Lebensqualität. Online im Internet.
Adresse: http://ec.europa.eu/research/quality-of-life/leaflets/de/casestud02.html.

[124] Internetquelle (04/2008):
Die Homepage des ExPASy Proteomics Server. Online im Internet.
Adresse: http://www.expasy.ch.

[125] Lehninger, A. (2001):
Lehninger Biochemie.
3. Aufl., Berlin, Heidelberg, New York: Springer Lehrbuch.

[126] Dressler, D. & Huntington, P. (1992):
Katalysatoren des Lebens.
Spektrum Bibliothek, Band 33, Heidelberg, Berlin, New York: Sepktrum Akademischer Verlag, Heidelberg.

[127] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2003):
Molekularbiologie der Zelle.
4. Edition, Weinheim: Wiley-VCH Verlag GmbH.

[128] Internetquelle (04/2008):
Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Online im Internet.
Adresse: http://www.chem.qmul.ac.uk/iubmb/enzyme.

[129] Patterson, C. (1988):
Homology in classical and molecular biology.
Mol. Biol. Evol., 5(6):603-625. Review.

[130] Wesselink, J. (2003):
Pattern Discovery in Biomoleclular Sequences: A Case Study in Yeast.
Veröffentlicht im Internet.
Adresse: http://genome.imim.es/~jjw/thesis.pdf.

[131] Needleman, S.B. & Wunsch, C.D. (1970):
A general method applicable to the search for similarity in the amino acid sequence of two proteins.
J. Mol. Biol., 48(3):443-453.

[132] Internetquelle (04/2008): (1981):
Die Datenbank KEGG. Online im Internet.
Adresse: http://www.genome.jp/kegg.

[133] Kawaji, H., Yamaguchi, Y., Matsuda, H. & Hashimoto, A. (2001):
A graph-based clustering method for a large set of sequences using a graph partitioning algorithm.
Genome Inform. Ser. Workshop Genome Inform., 12:93-102.

[134] Saitou, N. & Nei, M. (1987):
The neighbor-joining method: a new method for reconstructing phylogenetic trees.
Mol. Biol. Evol., 4(4):406-425.

[135] Bairoch, A. & Boeckmann, B. (1994*):*
The SWISS-PROT protein sequence data bank: current status.
Nucleic Acids Res., 22(17):3578-3580.

[136] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003):
The SWISS-PROT protein knowledgebase and its supplement TrEMBLE in 2003.
Nucleic Acids Res., 31(1):365-370.

[137] Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K. & Bairoch, A. (2002):
The PROSITE database, its status in 2002.
Nucleic Acids Res., 30(1):235-238.

[138] Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. & Bairoch, A. (2004*):*
Recent improvements to the PROSITE database.
Nucleic Acids Res., 32 Database issue:D134-D137.

[139] Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. & Zardecki, C.(2002):
The protein Data Bank.
Acta. Crystallogr. D. Biol. Crystallogr., 58(Pt6No1):899-907.

[140] Schomburg, I., Chang, A. & Schomburg, D. (2002):
BRENDA, enzyme data and metabolic information.
Nucleic Acids Res., 30(1):47-49.

[141] Internetquelle (04/2008):
Die Programmiersprache Python. Online im Internet.
Adresse: http://www.python.org.

[142] Internetquelle (04/2008):
Die Homepage von PyMOL. Online im Internet.
Adresse: http://pymol.sourceforge.net.

[143] Internetquelle (04/2008):
Die Homepage von yWorks Products. Online im Internet.
Adresse: http://www.yworks.com/en/products.htm.

[144] Altamarino, M.M., Plumbridge, J.A., Barba, H.A. & Calcagno, M.L. (1993):
Glucosamine-6-phosphate deaminase from Escherichia coli has a trimer of dimers structure with three intersubunit disulphides.
Biochem. J., 295 (Pt3):645-648.

[145] Monod, J., Wyman, J. & Changeux, J.P. (1965):
On the nature of allosteric transitions: a plausible model.
J. Mol. Biol., 12:88-118.

[146] Oliva, G., Fontes, M.R., Garratt, R.C., Altamirano, M.M., Calcagno, M.L. & Horjales, E. (1995):
Structure and catalytic mechanism of glucosamine 6-phosphate deaminase from Escherichia coli at 2.1 A resolution.
Structure, (12):1323-1332.

[147] Altamirano, M.M., Plumbridge, J.A., Horjales, E. & Calcagno, M.L. (1995):
Asymmetric allosteric activation of Escherichia coli glucosamine-6-phosphate deaminase produced by replacements of Tyr 121.
Biochemistry, 34(18):6074-6082.

[148] Midelfort, C.F. & Rose, I.A. (1977):
Studies on the mechanism of Escherichia coli glucosamine-6-phosphate isomerase.
Biochemistry, 16(8):1590-1596.

[149] Montero-Moran, G.M., Lara-Gonzales, S., Alvarez-Anorve, L.I., Plumbridge, J.A. & Calcagno, M.L. (2001):
On the multiple functional roles of the acitve site histidine in catalysis and allosteric regulation of Escherichia coli glucosamine 6-phosphate deaminase.
Biochemistry, 40(34):10187-10196.

[150] Cisnero, D.A., Montero-Moran, G.M., Lara-Gonzales, S. & Calcagno, M.L. (2004):
Inversion of the allosteric response of Escherichia coli glucosamine-6-P deaminase to N-acetylglucosamine 6-P, by single amino acid replacements.
Arch. Biochem. Biophys., 421(1):77-84.

[151] Lara-Gonzales, S., Dixon, H.B., Mendoza-Hernandez, G., Altamirano, M.M. & Calcagno, M.L. (2000):
On the role of the N-terminal group in the allosteric function of glucosamine-6-phosphate deaminase from Escherichia coli.
J. Mol. Biol., 301(1):219-227.

[152] Montero-Moran, G.M., Horjales, E., Calcagno, M.L. & Altamirano, M.M. (1998):
Tyr 254 hydroxyl group acts as a two-way switch mechanism in the coupling of heterotropic and homotropic effects in Escherichia coli glucosamine- 6-phosphate deaminase.
Biochemistry, 37(21):7844-7849.

[153] Joerger, A.C., Mueller-Diekmann, C. & Schulz, G.E. (2000):
Structures of L-Fuculose-1-Phosphat Aldolase mutants outlining motions during catalysis.
J. Mol. Biol., 303(4): 531-543.

[154] Joerger, A.C., Gosse, C., Fessner, W.D. & Schulz, G.E. (2000):
Catalytic action of fuculose 1-phosphate aldolase (class II) as derived from structure-directed mutagenesis.
Biochemistry, 39(20):6033-6041.

[155] Luo, Y., Samuel, J., Mosimann, S.C., Lee, J.E., Tanner, M.E. & Strynadka, N.C. (2001):
The structure of L-ribulose-5-phosphate 4-epimerase: an aldose-like platform for epimerization.
Biochemistry, 40(49):14763-14771.

[156] Johnson, A.E. & Tanner, M.E. (1998):
Epimerization via carbon-carbon bond cleavage. L-ribulose-5-phosphate 4-epimerase as a masked class II aldolase.
Biochemistry, 37(16): 5746-5754.

[157] Samuel, J., Luo, Y., Morgan, P.M., Strynadka, N.C. & Tanner, M.E. (2001):
Catalysis and binding in L-ribulose-5-phosphate 4-epimerase: a comparison with L-fuculose-1-phosphate aldolase.
Biochemistry, 40(49):14772-14780.

[158] Lee, L.V., Vu, M.V. & Cleland, W.W. (2000):
^{13}C and deuterium isotope effects suggest an aldol cleavage mechanism for L-ribulose-5-phosphate 4-epimerase.
Biochemistry, 39(16):4808-4820.

[159] Thoden, J.B., Miran, S.G., Philips, J.C., Howard, A.J., Raushel, F.M. & Holden, H.M. (1998):
Carbamoyl phosphate synthetase: caught in the act of glutamine hydrolysis.
Biochemistry, 37(25):8825-8831.

[160] Thoden, J.B., Raushel, F.M., Benning, M.M., Rayment, I. & Holden, H.M. (1999):
The structure of carbamoyl phosphate synthetase determined to 2.1 A resolution.
Acta. Crystallogr. D. Biol. Crystallogr., 55(Pt1):8-24.

[161] Miles, E.W., Rhee, S. & Davies, D.R. (1999):
The molecular basis of substrate channeling.
J. Biol. Chem., 274(18):12193-12196.

[162] Thoden, J.B., Huang, X., Raushel, F.M. & Holden, H.M. (1999):
The small subunit of carbamoyl phosphate synthetase: snapshots along the reaction pathway.
Biochemistry, 38(49):16158-16166.

[163] Huang, X. & Raushel, F.M. (1999):
Deconstruction of the catalytic array within the amidotransferase subunit of carbamoyl phosphate synthetase.
Biochemistry, 38(48):15909-15914.

[164] Akutsu, T. (2004):
Efficient extraction of mapping rules of atoms from enzymatic reaction data.
J Comput Biol., 11(2-3):449-62.

[165] Cerruela García, G., Luque Ruiz, I. & Gómez-Nieto, M.A. (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.

[166] Raymond, J.W. & Willett, P. (2002):
Maximum common subgraph isomorphism algorithms for the matching of chemical structures.
J Comput Aided Mol Des., 16(7):521-33.

[167] Gardiner, E.J., Artymiuk, P.J. & Willett, P. (1997):
Clique-detection algorithms for matching three-dimensional molecular structures.
J Mol Graph Model., 15(4):245-53.

CLUSTAL W Einstellungen

Einstellungen, soweit sie die Arbeitsweise des Programms beeinflussen:

2 ☞Multiple Alignments

2.4 ☞Toggle Slow/Fast pairwaise alignments = SLOW

2.5
2.5.1 ☞Gap Open Penalty: 10,00
2.5.2 ☞Gap Extension Penalty: 0,10
2.5.3 ☞Protein weight matrix: Gonnet series
2.5.4 ☞DNA weight matrix: IUB
2.5.5 ☞Gap penalty: 3
2.5.6 ☞K-tuple (word) size: 1
2.5.7 ☞No. of top diagonals: 5
2.5.8 ☞Window size: 5
2.5.9 ☞Toggle Slow/Fast pairwise alignments = SLOW

2.6
2.6.1 ☞Gap Opening Penalty: 10,00
2.6.2 ☞Gap Extension Penalty: 0,20
2.6.3 ☞Delay divergent sequences: 30%
2.6.4 ☞DNA Transitions Weight: 0,50
2.6.5 ☞Protein weight matrix: Gonnet series
2.6.6 ☞DNA weight matrix: IUB
2.6.7 ☞Use negativ matrix: OFF

2.6.8
2.6.8.1 ☞Toggle Residue-Specific Penalty: ON
2.6.8.2 ☞Toggle Hydrophilic Penalties: ON
2.6.8.3 ☞Hydrophilic Residues: GPSNDQEKR
2.6.8.4 ☞Gap Separation Distance: 4
2.6.8.5 ☞Toggle End Gap Separation:OFF
2.7 ☞Reset Gaps before alignment?: OFF

Übersicht über die Tabellen der Datenbanken *tee*

Der Aufbau der Tabelle *tree:*

Spalte	**Erklärung**
auto	auto_increment
tree_no	Nummer des Clusterbaums
evalue	Expectation Value, E-Value, E-Wert
Cluster	Clusternummer
node	Knoten
xco	x-Koordinate für die Darstellung des Clusterbaums
yco	y-Koordinate für die Darstellung des Clusterbaums
d_base	Datenbank, aus der Clustersequenzen stammen
align	Markierung für ein Alignment
sequenz_anz	Anzahl der Sequenzen im Cluster
n_hicon	Anzahl hochkonservierter Positionen im Sequenzmuster
n_con	Anzahl konservierter Positionen im Sequenzmuster
pat_len	Länge des Sequenzmusters
pattern	Vollständiges Sequenzmuster

Der Aufbau der Tabelle *edges:*

Spalte	**Erklärung**
auto	auto_increment
tree_no	Nummer des Clusterbaums
source	Ausgangscluster für die Darstellung des Clusterbaums
target	Zielcluster für die Darstellung des Clusterbaums

Der Aufbau der Tabelle *ecnumbers:*

Spalte	**Erklärung**
auto	auto_increment
tree_no	Nummer des Clusterbaums
node	Knoten
num	Anzahl Sequenzen mit jeweiliger EC-Nummer
ec	EC-Nummer

Der Aufbau der Tabelle *veri:*

Spalte	**Erklärung**
auto	auto_increment
tree_no	Nummer des Clusterbaums
node	Knoten
evalue	Expectation Value, E-Value, E-Wert
d_base	Datenbank, aus der Clustersequenzen stammen
ec	EC-Nummer(n) des Sequenzmusters
sequenz_anz	Anzahl der Sequenzen im Cluster
pat_len	Länge des Sequenzmusters
rp	Anzahl erzielter Richtig-Positiver Treffer
fp	Anzahl erzielter Falsch-Positiver Treffer

Der Aufbau der Tabelle *hits:*

Spalte	**Erklärung**
count	auto_increment
auto	Link zur Datenbank tree
sequenz_anz	Anzahl der Sequenzen im Cluster
d_base	Datenbank, aus der Clustersequenzen stammen
hit	Uniprot-Number der getroffenen Sequenz
d_base_hit	Datenbank, aus der die getroffenen Sequenzen stammen

Der Aufbau der Tabelle *xt:*

Spalte	**Erklärung**
region_id	auto_increment
sequence_id	Identifizierung der Sequenz
region_start	Start der Domänenregion
region_end	Ende der Domänenregion
xe	Clusterzusammensetzung bei E-Wert x

Verzeichnisstruktur der DVD

1. Verzeichnis Daten:

1.1 uniprot_sprot.rar
1.2 uniprot_trembl.rar

2. Verzeichnis Datenbanken:

2.1 tee_final.rar
2.2 seq.rar
2.3 regiondb.sql.gz

3. Verzeichnis Dissertation:

3.1 Dissertation.pdf

4. Verzeichnis Programme:

4.1 clustalw-1.83-2.i586.rpm
4.2 yed3_0_0_1.exe
4.3 pymol-0_97-bin-win32.zip
4.4 wrar371d.exe
4.5 Verzeichnis: Eigene Programme

5. Verzeichnis Beispiele:

5.1 Clusterbaum Beispiel 1
5.2 Clusterbaum Beispiel 2
5.3 Clusterbaum Beispiel 3

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 dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. D. Schomburg betreut worden.

Unterschrift Adrian Welfle