Identifizierung ähnlicher Reaktionsmechanismen in homologen Enzymen unterschiedlicher Funktion unter Verwendung konservierter Sequenzdomänen

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

Christian aus dem Spring

aus Krefeld

Köln 2006

Berichterstatter/in: Prof. Dr. D. Schomburg

Prof. Dr. S. Waffenschmidt

Tag der letzten mündlichen Prüfung:

Wertlos, was ohne Anstrengung entworfen.

Michelangelo

Danksagung

An dieser Stelle möchte ich allen danken, die zu dieser Arbeit beigetragen haben.

Mein besonderer Dank gilt meinem Doktorvater Prof. Dr. D. Schomburg für die interessante Themenstellung, seine ständige Bereitschaft zum fachlichen Dialog sowie den großzügig gewährten Freiraum bei der Ausgestaltung dieser Arbeit. Darüber hinaus möchte ich mich bei Ihm für das mir entgegengebrachte Vertrauen sowie seine anhaltende Unterstützung bedanken.

Frau Prof. Dr. S. Waffenschmidt danke ich für die freundliche Übernahme des Koreferates.

Darüber hinaus danke ich allen Mitarbeitern der Arbeitsgruppe Schomburg für die angenehme Arbeitsatmosphäre und ihre Hilfsbereitschaft. Besonderns hervorheben möchte ich in diesem Zusammenhang Frau Dr. Antje Chang und Frau Dr. Ida Schomburg, die mir stets unterstützend zur Seite standen und maßgebend zum Gelingen dieser Arbeit beigetragen haben. Darüber hinaus danke ich Herrn Dr. Oliver Hofmann, der mir bei Fragen der Programmiertechnik und der konzeptionellen Gestaltung eine große Hilfe war.

Außerdem möchte ich mich ganz herzlich bei Herrn Dr. Dominik Stuhlmann für die Durchsicht meiner Arbeit und die moralische Unterstützung bedanken.

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, Diagrammen 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 der in Anhang D angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Dietmar Schomburg betreut worden.

Christian aus dem Spring

Inhaltsverzeichnis

Abkürzungsverzeichnis

BLAST	Basic Local Alignment Search Tool
BLASTP	BLAST für Proteinsequenzen
BLOSUM	Blocks Substitution Matrix
CATH	Class(C), Architecture(A), Topology(T) and Homologous superfamily (H) Protein Structure Classification
CAZy	Carbohydrate-Active Enzymes
DDD	Dali Domain Dictionary
DE	Description Field
DHAP	Dihydroxyacetonphosphat
DNA	Deoxy Nucleic Acid
EC	Enzyme Commission
EMBL	European Molecular Biology Laboratory
E-value	Expectation Value
et al.	et aliter
EXPASY	Expert Protein Analysis System
FASTA	Fast Alignment
FTP	File Transfer Protokoll
FucA	L-Fuculose-phosphat-Aldolase
HSP	High-scoring Segment Pair
ID	Identification
IUB	International Union of Biochemistry
IUBMB	International Union of Biochemistry and Molecular Biology
LCR	Low Complexity Region
MSA	Multiple Sequence Alignment
NADH	reduziertes Nicotinsäureamidadenindinucleotid
NADPH	reduziertes Nicotinsäureamidadenindinucleotid-phosphat
NMR	Nuclear Magnetic Resonance
NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank

PIR .....Protein Information Resource
RhuA Rhamnulose-1-phosphat-Aldolase
RibE .....L-Ribulose-phosphat 4-Epimerase
SCOP.....Structural Classification of Proteins
SIB .....Swiss Institute of Bioinformatics
TrEMBL.....Translated EMBL

Kurzzusammenfassung

Enzyme sind außerordentlich effiziente Biokatalysatoren und beschleunigen als solche nahezu sämtliche biochemischen Reaktionen in biologischen Systemen. Neue Enzyme entstehen nicht *de novo*, sondern entwickeln sich schrittweise durch Abwandlung der bereits vorhandenen Enzyme. Daher lassen sich die Reaktionen des Grundstoffwechsels der Zellen trotz ihrer Vielfalt auf relativ wenige Grundtypen zurückführen. Diese Tatsache hat man teilweise bei der EC-Klassifikation der Enzyme berücksichtigt. Die Einordnung in EC-Klassen erfolgt jedoch im allgemeinen nicht aufgrund von gemeinsamer Abstammung oder ähnlichen Reaktionsmechanismen, sondern überwiegend nach enzymologischen Kriterien wie der Wirkungs- und Substratspezifität. Infolgedessen weisen Enzyme der gleichen EC-Klasse häufig keine strukturelle Ähnlichkeit zueinander auf, wodurch impliziert wird, daß diese Enzyme eher durch Konvergenz als durch Divergenz entstanden sind, während umgekehrt Enzyme gemeinsamen evolutionären Ursprungs oftmals ganz unterschiedlichen EC-Klassen angehören. Letzteres führte zur Annahme, daß Enzyme trotz gemeinsamer Abstammung ganz verschiedene Funktionen haben können. Es gibt jedoch Hinweise darauf, daß diese Enzyme ähnliche Reaktionsmechanismen zur Realisierung der verschiedenen Funktionen verwenden. Während die EC-Klassifikation alle an sie gestellten Anforderungen erfüllt, besteht somit Bedarf für ein alternatives, komplementäres Klassifizierungssystem, das nicht auf einer empirischen Einteilung der beobachteten Reaktionen, sondern auf der evolutionären Verwandtschaft der Enzyme beruht und infolgedessen Rückschlüsse auf die zugrundeliegenden Reaktionsmechanismen zuläßt.

In der vorliegenden Dissertation wurde untersucht, ob eine auf Sequenzhomologie basierende Einteilung der Enzyme mit den von den Enzymen verwendeten Reaktionsmechanismen korreliert. Ziel war die systematische Clusterung und Analyse aller bekannten Enzymsequenzen zur Identifizierung von gemeinsamen oder ähnlichen Enzymmechanismen. Vorbedingung zur Bearbeitung des Problems war die Entwicklung einer Methode zur Identifizierung modular aufgebauter Proteinen, die aus mehreren, evolutionär oftmals unabhängigen Sequenzdomänen bestehen. Da solche modularen Enzyme in unterschiedlichen Bereichen Ähnlichkeit zu verschiedenen Enzymfamilien aufweisen können,

implizieren sie häufig ein scheinbares, tatsächlich jedoch nicht vorhandenes gemeinsames Auftreten von Enzymaktivitäten in einem Sequenzcluster.

Die Domänenstruktur wurde mittels der Lage und Ausdehnung lokaler Sequenzalignments ermittelt. Anschließend wurden die so bestimmten Sequenzbereiche entsprechend ihrer Sequenzähnlichkeit zu Gruppen homologer Sequenzabschnitte zusammengefaßt. Hierzu wurde die Methode der Clusteranalyse verwendet. Die Analyse erfolgte bei verschiedenen Grenzwerten, um eine hierarchische Strukturierung des Sequenz-Raumes zu erhalten. Hierbei zeigte sich, daß abhängig vom verwendeten Grenzwert bis zu 40% der generierten Sequenzcluster Enzyme verschiedener Enzymklassen, teilweise sogar verschiedener EC-Hauptklassen enthielten. Bei der Analyse zeigte sich jedoch, daß in allen betrachteten Fällen trotz auf den ersten Blick unterschiedlicher Katalyse der Reaktionsmechanismus oder aber die Substratspezifität dieser Reaktionen sehr ähnlich sind.

Abstract

Enzymes are highly efficient biocatalysts that accelerate almost all biochemical reactions in biological systems. New enzymes do not arise *de novo* but have developed in a stepwise manner from pre-existing forms. Therefore the main reactions in cellular metabolism, despite their great variety, can be grouped into a small number of basic reaction types. This phenomenon has partially been taken account of in the EC system for enzyme classification. The allocation to EC classes, however, is not based on common ancestry or similar reaction mechanisms but rather on substrate specificity. Hence the enzymes in a class often fail to show structural similarity to each other; this implies that they have developed convergently and not divergently, whereas enzymes of common origin often belong to different EC classes. The latter has led to the assumption that enzymes, despite their common origin, may have completely different functions. However these enzymes might apply similar catalytic mechanisms for their various functions. Whilst the EC classification does everything it was intended to do, an alternative complementary concept based on the evolutionary relationships of the enzymes, and not on the empirical classification of reactions, is required in order to allow conclusions to be drawn on the underlying reaction mechanisms.

This work examines whether a correlation can be made between a grouping of enzymes based on their sequence homology and their catalytic mechanisms. The aim has been systematically to cluster and analyse all known enzyme sequences in order to identify common or similar enzyme mechanisms. A prerequisite has been the development of a method for the identification of proteins, which are composed of multiple sequence domains that have arisen independently of evolution. As such enzymes may show similarities with diverse domains of other enzyme families; this implies that they exhibit an apparent, but illusionary, common enzyme activity in one sequence cluster.

The structure of the domains has been determined by the position and extention of local sequence alignments. Subsequently, the sequence regions obtained by this method have been classified with regard to their sequence similarity to groups of homologous protein sections by using cluster analysis. This analysis has been carried out by applying different boundary conditions in order to ob-

tain a hierarchical structure of the sequence space. Depending on the conditions, up to 40% of such generated clusters contain enzymes from different enzyme classes, some even from different major EC classes. The analysis, however, has shown that, for all cases examined, the reaction mechanism or the substrate specificity is similar, irrespective of the different function.

1 Einleitung

1.1 Enzyme

Enzyme sind außerordentlich effiziente Biokatalysatoren und beschleunigen als solche nahezu sämtliche chemischen Reaktionen in biologischen Systemen. Ohne Katalysatoren laufen die meisten chemischen Umsetzungen unter physiologischen Bedingungen nur in geringem Umfang ab. Enzyme bewirken jedoch eine Beschleunigung der Reaktionsgeschwindigkeit um mindestens den Faktor 10^6 [1]. Enzyme sind überdies an vielen Regulationsmechanismen beteiligt, die es ermöglichen, den Stoffwechsel an veränderte Bedingungen anzupassen. Tatsächlich ist der organisierte Ablauf der zahlreichen chemischen Prozesse nur möglich, weil jede Zelle über eine eigene, genetisch determinierte Enzymausstattung verfügt. Erst dadurch kommt es zu koordinierten Reaktionsfolgen, die in ihrer Summe den Stoffwechsel eines Organismus ausmachen.

1.1.1 EC-Klassifikation der Enzyme

Zu Beginn der Enzymologie wurden Enzyme – meist nach Belieben der Entdekker – mit Trivialnamen benannt. Da es keine einheitliche Regelung zur Benennung von Enzymen gab, kam es gelegentlich vor, daß mehrere verschiedene Bezeichnungen für dasselbe Enzym verwendet wurden, während umgekehrt verschiedene Enzyme denselben Namen erhielten. Überdies ließen viele Enzymnamen, wie zum Beispiel „Katalase“, keinen Rückschluß auf die katalysierte Reaktion zu [2].

Zur Verbesserung der Situation und Aufstellung von Regeln für eine rationelle Namensgebung der schnell wachsenden Anzahl neu entdeckter Enzyme, etablierte die Enzymkommission (Enzyme Commission = EC) der *International Union of Biochemistry* (IUB; heute IUBMB) im Jahre 1964 eine Schema für die systematische funktionelle Klassifizierung der Enzyme. Dieses als EC-Klassifikation bezeichnete Schema klassifiziert Enzyme entsprechend der Natur der von ihnen katalysierten chemischen Reaktionen [3].

Jedes Enzym wird unter zwei Namen und einer vierstelligen EC-Nummer im Enzym-Katalog geführt. Der empfohlene Name entspricht häufig dem früher benutzten Trivialnamen. Der systematische Name eines Enzyms wird verwen-

det, um Zweideutigkeiten zu vermeiden und wird aus der Bezeichnung des Substrates, des beteiligten Co-Enzyms (Co-Substrates) und des Reaktionstyps des Enzyms abgeleitet. Die vierstellige EC-Nummer ergibt sich wie folgt:

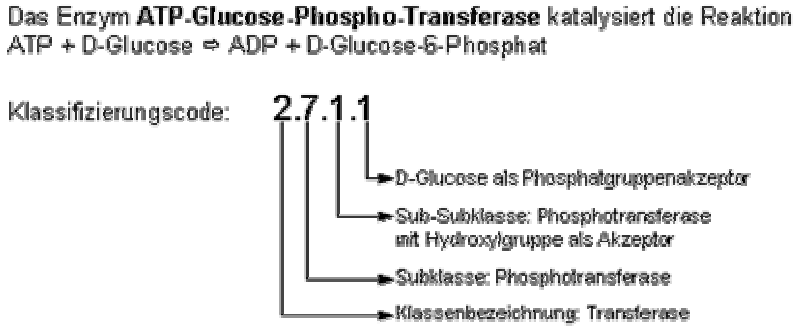


Abbildung 1-1: Beispiel für die Klassifizierung des EC-Systems.

Die erste Ziffer gibt die Zugehörigkeit zu einer der sechs Hauptklassen an, in denen jeweils Enzyme ähnlicher Wirkungsspezifität zusammengefaßt sind (vgl. Tabelle 1-1). Die nächsten beiden Ziffern definieren Sub- bzw. Sub-Subklassen und ergeben sich aus dem umgesetzten Substrat, dem beteiligten Co-Enzym oder Akzeptor und weiteren Kriterien. Die letzte Ziffer ist schließlich die laufende Nummer des Enzyms innerhalb seiner Sub-Subklasse.

Die Einordnung in die EC-Klassifikation erfolgt aufgrund der beobachteten Wirkungs- und Substratspezifität der Enzyme. Ihre evolutionäre Beziehung oder der zugrundeliegende Reaktionsmechanismus werden dagegen nicht berücksichtigt. Dies führt dazu, daß nichtverwandte Enzyme, welche die gleiche Reaktion mittels unterschiedlicher Mechanismen realisieren, die gleiche EC-Klassifikation erhalten, während Enzyme gemeinsamer Abstammung, die ähnliche Mechanismen zur Katalyse unterschiedlicher Reaktionen verwenden, verschiedenen EC-Klassen zugeordnet werden.

Während die EC-Klassifikation alle an sie gestellten Anforderungen erfüllt, besteht somit Bedarf für ein alternatives, komplementäres Klassifizierungssystem der Enzyme, das nicht auf einer empirischen Einteilung der beobachteten Reaktionen, sondern auf der evolutionären Verwandtschaft der Enzyme beruht und infolgedessen Rückschlüsse auf den zugrundeliegenden Reaktionsmechanismus zuläßt.

Tabelle 1-1: Einteilung der Enzyme gemäß EC-Klassifikation. Enzyme werden entsprechend ihrer katalytischen Funktion in sechs Hauptklassen eingeteilt.

Enzym-Klasse		Subklassen	Reaktionstyp
1	**Oxidoreduktasen**	Dehydrogenasen Oxidasen	Elektronentransfer zwischen zwei Substraten bzw. einem Substrat und einem Co-Enzym (Co-Substrat)
2	**Transferasen**	Kinasen Acetyl-, Methyl-, Aminotransferasen Polymerasen	Transfer von Gruppen von einem Donor (oft ein Co-Enzym) auf einen Akzeptor (meist das Substrat)
3	**Hydrolasen**	Proteasen Esterasen Nukleasen Glucosidasen ATPasen	hydrolytische Spaltung
4	**Lyasen**	Decarboxylasen Aldehydlyasen Hydrolyasen Synthasen	Abspaltung von Gruppen nach einem nicht-hydrolytischen Mechanismus
5	**Isomerasen**	Racemasen Epimerasen Isomerasen	intramolekularer Transfer von Gruppen, ohne daß sich die Bruttoformel des Substrates ändert
6	**Ligasen**	Synthetasen	Bildung von C-C-, C-N-, C-O- oder C-S-Bindungen durch Kondensation unter Spaltung energiereicher Phosphatbindungen

1.2 Analyse von Proteinsequenzen

Proteine sind trotz ihrer funktionellen Vielseitigkeit eine relativ homogene Klasse von Molekülen. Sie alle sind lineare, unverzweigte Polymere, die sich aus verschiedenen Kombinationen der gleichen 20 proteinogenen Aminosäuren zusammensetzen [4]. Sie unterscheiden sich im wesentlichen nur in der Anzahl sowie der Abfolge, mit der die einzelnen Aminosäuren zu polymeren Ketten verknüpft sind.

Die funktionelle Vielseitigkeit der Proteine ist zum Teil auf die unterschiedlichen chemischen Eigenschaften der verschiedenen Aminosäuren zurückzuführen. Sie beruht jedoch vor allem auf der spezifischen dreidimensionalen

Struktur, die sich - in Abhängigkeit von der jeweiligen Aminosäuresequenz - durch die Faltung der linearen Polypeptidkette ergibt. So ergaben Denaturierungs- und anschließende Renaturierungsversuche, daß die Aminosäuresequenz (Primärstruktur) indirekt auch die höheren Ebenen der Proteinstruktur (Sekundär-, Tertiär- und Quartärstruktur) determiniert [5]. Die Aminosäuresequenz bestimmt somit sowohl die strukturellen als auch die funktionellen Eigenschaften eines Proteins. Daher ist die Analyse der Sequenz ein grundlegender Schritt bei der Charakterisierung eines Proteins.

1.2.1 Strukturelle und funktionelle Verwandtschaft homologer Proteine

Eine isolierte Betrachtung der Primärstruktur vermag nur wenig Aufschluß über die strukturellen und funktionellen Eigenschaften eines Proteins zu geben. Die Aminosäuresequenz kann jedoch zur Charakterisierung eines Proteins beitragen, wenn die Sequenz homolog zu der eines anderen Proteins bekannter Struktur und Funktion ist, da homologe Proteine die gleiche räumliche Struktur und häufig auch identische oder verwandte Funktionen besitzen [6, 7].

Homologe Proteine gehen auf einen gemeinsamen „Vorfahr“ zurück, aus dem sie durch divergente Evolution hervorgegangen sind [8]. Im Laufe der Evolution akkumulierten die codierenden Gene jedoch Mutationen und brachten so Variationen der Proteine hervor, die zwar häufig noch die gleiche biologische Funktion erfüllen, sich aber mehr oder minder in ihrer Primärstruktur unterscheiden. Das Ausmaß der Sequenzunterschiede hängt einerseits davon ab, inwieweit die Mutationen die Proteinfunktion beeinflussen, da Mutationen, die eine Beeinträchtigung oder gar den Verlust der Funktion zur Folge haben, meist der natürlichen Selektion unterliegen und andererseits davon, wie lange die evolutionäre Divergenz der codierenden Gene bereits zurückliegt. Die Sequenzen homologer Proteine können daher nahezu identisch, bis zu einem gewissen Grad ähnlich, oder aufgrund weitreichender Mutationen bis auf wenige essentielle Positionen vollkommen verschieden sein.

Bei relativ nah verwandten Proteinen kann Homologie über einen Sequenzvergleich aufgedeckt werden. Zeigen zwei Sequenzen eine signifikante Sequenzähnlichkeit zueinander, kann nahezu sicher von Homologie ausgegangen werden, denn angesichts der großen Anzahl theoretisch möglicher Protein-

sequenzen[1] erscheint es unwahrscheinlich, daß sich eine signifikante Sequenzähnlichkeit zufällig oder unter dem Selektionsdruck für die gleiche Funktion aus evolutionär unabhängigen Sequenzen entwickelt hat (konvergente Evolution). Hohe Sequenzähnlichkeit ist daher stets ein Indikator für die Verwandtschaft der Sequenzen [9].

1.2.2 Sequenzalignment als Methode zur Identifikation homologer Proteine

Um den Grad der Ähnlichkeit zwischen zwei oder mehr Proteinsequenzen zu ermitteln, verwendet man die Methode des Sequenzalignments. Es stehen eine Reihe verschiedener Algorithmen und Programme zur Verfügung, um Sequenzen miteinander zu vergleichen. Die ersten und noch immer verwendeten Algorithmen zur Berechnung eines Alignments stammen aus den frühen siebziger und achtziger Jahren des 20. Jahrhundert und basieren auf einem dynamischen Programmieralgorithmus [10, 11]. Sie sind darauf ausgerichtet, ein möglichst optimales Alignment der Sequenzen zu berechnen. Ist das Ziel eines Sequenzvergleichs weniger das eigentliche Alignment, sondern die Identifikation homologer Sequenzbereiche, können geschwindigkeitsoptimierte Verfahren wie FASTA [12] und BLAST [13] verwendet werden. Die biologische Signifikanz der Sequenzähnlichkeit ist statistisch gut untersucht [14, 15].

1.2.2.1 Statistische Signifikanz eines Alignments als Indikator für Homologie

Beim Alignmentverfahren wird ein Maß für die Ähnlichkeit zweier Sequenzen ermittelt, welches den Grad der Übereinstimmung zwischen den beiden Sequenzen angibt. Wenn die Sequenzen über weite Bereiche identisch sind, kann nahezu sicher davon ausgegangen werden, daß die Proteine homolog zueinander sind und infolgedessen eine ähnliche Faltung und Funktion besitzen. Die Interpretation schwacher Sequenzidentität ist hingegen nicht eindeutig.

---

[1] Wenn man eine durchschnittliche Sequenzlänge von 300 Aminosäuren zugrundelegt, so ergeben sich bei 20 verschiedenen Aminosäuren, die jede der 300 Positionen einnehmen können, 20^{300}, also insgesamt über 10^{390} Kombinationen für Proteinsequenzen dieser Länge. Die Zahl derzeit bekannter Sequenzen liegt indes bei etwa 10^{6}.

Es läßt sich nur schwer ein Grenzwert für die Sequenzidentität definieren, der erforderlich ist, um zweifelsfrei die Homologie zweier Sequenzen zu belegen, da dieser sowohl von der Länge als auch der Aminosäurezusammensetzung der Sequenzen abhängig ist [16]. Wenn alle 20 Aminosäuren mit derselben Häufigkeit vorkommen, dann erwartet man bei zufällig gewählten Sequenzen, die nicht miteinander verwandt sind, eine Übereinstimmung in 5% der Aminosäurereste [17]. Für lange Sequenzen mit einer typischen Aminosäurezusammensetzung wird im allgemeinen eine Übereinstimmung in mehr als 30% der Aminosäurereste als wahrscheinlicher Beleg für eine ähnliche Struktur [16, 18, 19] und Funktion [7] angesehen .Der Bereich von 15-30% Sequenzidentität wird indes als "*twilight zone*" bezeichnet [20], da in diesem Bereich oftmals keine klare Aussage über eine Homologie der Sequenzen getroffen werden kann. Es sind jedoch Proteine bekannt, die bei gleicher Faltung nur noch eine sehr geringe Sequenzähnlichkeit besitzen [6]. Diese Fälle können durch einfache Sequenzvergleiche nicht erfaßt werden.

1.2.3 Identifizierung entfernt homologer Proteine

Homologie ist eine transitive Beziehung. Wenn Sequenz A zu Sequenz B homolog ist und Sequenz B zu Sequenz C, dann besteht zwangsläufig auch eine evolutionäre Verwandtschaft zwischen den Sequenzen A und C, auch wenn diese keine signifikante Sequenzähnlichkeit mehr zueinander aufweisen sollten [7, 9].

Die Transitivität der Homologie kann genutzt werden, um verwandte Sequenzen über ein Maß der direkten Sequenzidentität hinaus zu ermitteln. Der Nutzen der Transitivität, um Homologie zwischen entfernt verwandten Proteinen abzuleiten, wurde bereits in verschiedenen Arbeiten demonstriert [6, 7, 21-25]. Eine auf diesem Wege nachgewiesene Homologie kann beispielsweise durch Strukturähnlichkeit bestätigt werden.

Die dreidimensionale Struktur von Proteinen ist evolutionär stärker konserviert als die Sequenz [16]. Die Ursache hierfür besteht darin, daß die Funktion eines Proteins im wesentlichen eine bestimmte Struktur erfordert, diese aber über verschiedene Sequenzen realisiert werden kann. Es ist daher möglich, durch einen Strukturvergleich entfernte evolutionäre Verwandtschaften festzustellen, die auf Sequenzebene nicht mehr nachzuweisen sind [26]. Da die experimentelle Bestimmung einer Proteinstruktur in der Regel jedoch weitaus

schwieriger und zeitaufwendiger als die Bestimmung einer Sequenz ist, stehen der Menge bekannter Proteinsequenzen nur vergleichsweise wenige aufgeklärte Strukturen gegenüber.

Eine weitere Möglichkeit, um schwach konservierte Homologien zwischen Proteinsequenzen zu belegen, ist die Verwendung charakteristischer Sequenzmotive und –profile. Diese Motive und Profile setzen sich aus invarianten und konservativ substituierten Aminosäurepositionen zusammen, die für die Ausübung einer Funktion meist von essentieller Bedeutung sind, und daher während der Evolution konserviert wurden. Solche Motive lassen sich häufig auch dann noch identifizieren, wenn die übrigen Bereiche der Sequenz bereits keine nachweisbare Ähnlichkeit mehr aufweisen.

1.2.4 Clusteranalyse als Methode zur Gruppierung homologer Proteine

Will man die Transitivität der Homologie nutzen, ist die Clusteranalyse eine geeignete Methode, um auch entfernt verwandte Sequenzen zu identifizieren. Die Clusteranalyse ist ein Verfahren zur Gruppenbildung und gehört zu den explorativen Methoden der Statistik. Sie befaßt sich mit dem Problem, homogene Teilmengen (Cluster) in einer insgesamt heterogenen Gesamtmenge von Objekten zu erkennen und zusammenzufassen, so daß die Mitglieder einer Gruppe einander in Bezug auf im voraus bestimmte Kriterien möglichst ähnlich sind (interne Homogenität), während zwischen den Gruppen eine möglichst geringe Ähnlichkeit besteht (externe Heterogenität). Ziel der Gruppierung ist eine Repräsentation der im Datensatz inhärent enthaltenen Struktur, um so Beziehungen zwischen den Objekten herauszustellen.

In der theoretischen Proteinforschung ist die Clusteranalyse eine häufig verwendete Methode und bildet die Grundlage für die Erstellung verschiedener Datenbanken. So bietet beispielsweise die PIR-Datenbank (Protein Information Resource) eine Klassifikation von Proteinen in Proteinfamilien anhand einer Sequenz-Clusterung [28]. Das Gruppieren von Sequenzen in Familien hat sich bereits als hilfreich für die Studie und das Verständnis der funktionellen sowie strukturellen Verwandtschaft von Proteinen erwiesen [29, 30]. So kann die mittels Clusteranalyse erhaltene Gruppierung genutzt werden, um

- den Sequenzraum zu strukturieren,

- die Datengrundlage für evolutionäre Untersuchungen zu erstellen und
- eine Hilfestellung für die Erstellung multipler Alignments zu bieten.

1.2.5 Proteindomänen als Bausteine modularer Proteinarchitektur

Sequenzvergleichsmethoden ermöglichen es, homologe Sequenzen zu identifizieren und zu gruppieren. Oftmals können Methoden des Sequenzvergleichs aber nur eine partielle Ähnlichkeit zwischen den Sequenzen zweier Proteine feststellen. Dies ist auf die modulare Struktur vieler Proteine zurückzuführen.

Proteine bestehen häufig aus Kombinationen mehrerer gleich- oder verschiedenartiger Substrukturen, oder Domänen [31, 32]. Diese werden daher auch als Mosaik- oder Multidomänenproteine bezeichnet. Manche der Proteindomänen können als evolutionär mobile Module angesehen werden, da sie in verschiedenen Proteinen in unterschiedlichem Domänenkontext auftreten [31, 33].

Die Wiederverwendung von Domänen spielt vor allem in der Evolution von Proteinen multizellulärer Organismen eine wichtige Rolle. Es wird vermutet, daß Multidomänenproteine durch einen als *domain shuffling* bezeichneten Prozeß entstehen [34]. Dabei werden vorhandene Module spezifischer Funktionen neu gruppiert, so daß ein neuartiges Protein mit neuen Eigenschaften entsteht, welche sich aus der Kombination der Einzeldomänen ableiten.

Da die aus mehreren evolutionär mobilen Domänen zusammengesetzten Proteine in unterschiedlichen Bereichen Ähnlichkeit zu verschiedenen Domänen anderer Proteinen aufweisen, stellt die Domänenstruktur ein erhebliches Erschwernis für die Klassifizierung der Proteinsequenzen dar. Aus diesem Grund ist die Kenntnis und entsprechenden Berücksichtigung der modularen Domänenarchitektur eine grundlegende Voraussetzung, um Sequenzen korrekt zu klassifizieren.

1.3 Zielsetzung

Biochemische Systeme sind trotz ihrer funktionellen Vielfalt stets auch ökonomische Systeme. Neue Enzyme und Stoffwechselwege entstehen nicht *de novo*, sondern entwickeln sich schrittweise durch Abwandlung der vorhandenen und bereits bewährten Systeme [35]. Daher lassen sich die Reaktionen des Grund-

stoffwechsels der Zellen trotz ihrer Vielfalt auf relativ wenige Grundtypen zurückführen. Diese Tatsache hat man teilweise bei der EC-Klassifikation der Enzyme berücksichtigt. Die Einordnung in EC-Klassen erfolgt jedoch im allgemeinen nicht aufgrund von gemeinsamer Abstammung oder ähnlichen Reaktionsmechanismen, sondern überwiegend nach enzymologischen Kriterien wie etwa der Wirkungsspezifität. Infolgedessen weisen Enzyme der gleichen EC-Klasse häufig keine strukturelle Ähnlichkeit zueinander auf, wodurch impliziert wird, daß diese Enzyme eher durch Konvergenz (Analogie) als durch Divergenz entstanden sind, während umgekehrt Enzyme gemeinsamen evolutionären Ursprungs oftmals ganz unterschiedlichen EC-Klassen angehören. Letzteres führte zur Annahme, daß Enzyme trotz evolutionärer Verwandtschaft ganz unterschiedliche Reaktionen katalysieren können. Es gibt jedoch Hinweise darauf, daß trotz auf den ersten Blick unterschiedlicher Katalyse der Reaktionsmechanismus dieser Reaktionen oftmals sehr ähnlich ist [36].

Die vorliegende Dissertation geht der Frage nach, ob eine auf Sequenzhomologie basierende Einteilung der Enzyme mit den von den Enzymen verwendeten Reaktionsmechanismen korreliert und somit zur mechanistischen Klassifizierung der Enzyme verwendet werden kann. Ziel ist die systematische Clusterung und Analyse aller bekannten Enzymsequenzen zur Identifizierung von gemeinsamen oder ähnlichen Enzymmechanismen.

Vorbedingung zur Bearbeitung des Problems war die Entwicklung einer Methode zur Identifizierung modular aufgebauter Proteine, die aus mehreren, evolutionär oftmals unabhängigen Sequenzdomänen bestehen. Da solche Multidomänensequenzen in unterschiedlichen Bereichen Ähnlichkeit zu verschiedenen Enzymfamilien aufweisen können, implizieren sie häufig ein scheinbares, tatsächlich jedoch nicht vorhandenes gemeinsames Auftreten von Enzymaktivitäten in einem Sequenz-Cluster.

2 Daten, Algorithmen und Methoden

2.1 Übersicht über den methodischen Verlauf der Arbeit

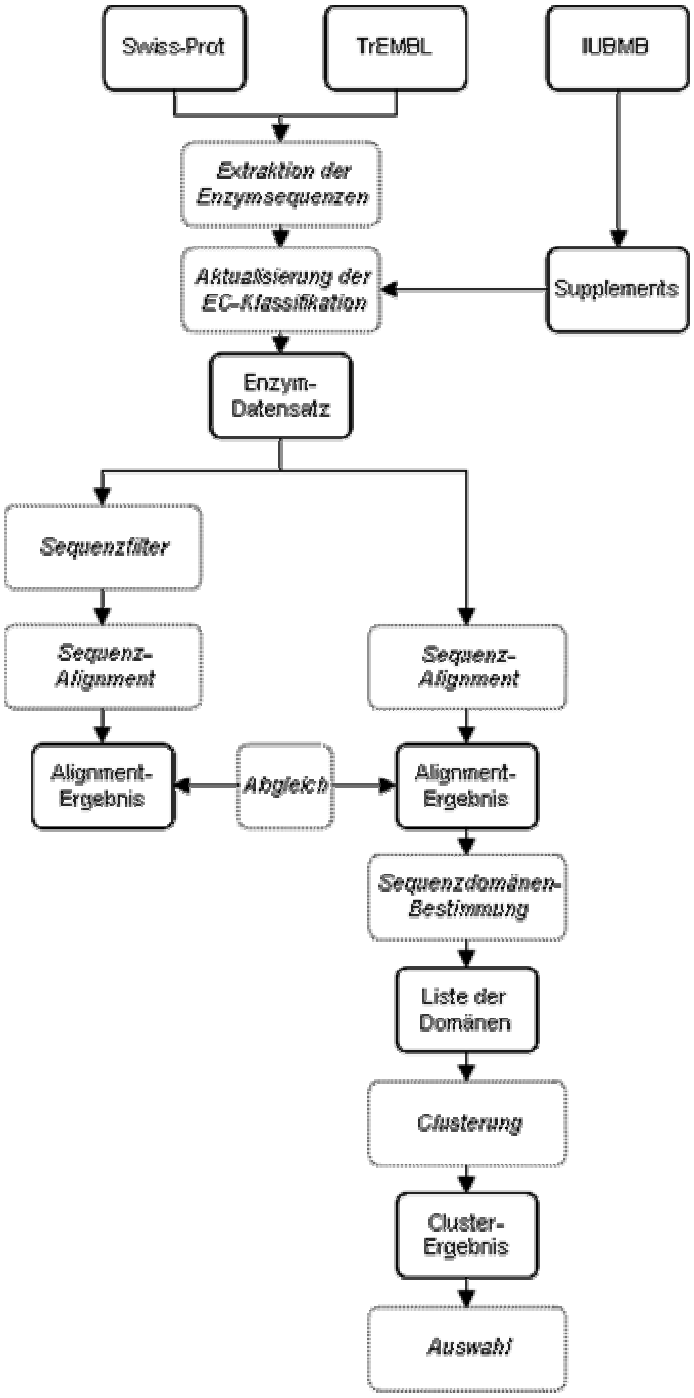


Abbildung 2-1: Schema des methodischen Gesamtverlaufs der Arbeit

In der vorliegenden Arbeit soll untersucht werden, ob Enzyme, die einen gemeinsamen phylogenetischen Ursprung haben, aber gemäß EC-System unterschiedliche Reaktionen katalysieren, ähnliche enzymatische Reaktionsmechanismen zur deren Realisierung verwenden. Zur Klärung dieser Frage wird folgende Strategie verfolgt (vgl. Abbildung 2-1):

Zunächst wird auf der Grundlage der Proteinsequenzdatenbanken Swiss-Prot und TrEMBL [37] ein Enzymdatensatz erstellt. Dieser enthält jene Sequenzeinträge aus beiden Datenbanken, die entsprechend des Enzymkataloges der IUBMB [38] klassifiziert sind und eine vollständige vierstellige EC-Nummer besitzen. Gleichzeitig werden alle EC-Nummern auf ihre Gültigkeit geprüft und gegebenenfalls nach den Vorgaben der IUBMB aktualisiert.

Um festzustellen, welche Enzyme eine signifikante Sequenzähnlichkeit zueinander aufweisen, die auf eine evolutionäre Verwandtschaft schließen läßt, werden alle Enzymsequenzen zweier vollständiger „alle vs. alle"-Sequenzalignments unterzogen, bei denen jede Sequenz paarweise mit allen übrigen des Datensatzes verglichen wird. Eines der „alle vs. alle"-Alignments erfolgt unter Verwendung der unveränderten Originalsequenzen, wie sie aus Swiss-Prot und TrEMBL extrahiert wurden. Für das andere werden die Enzymsequenzen zunächst mit einem Sequenzfilter prozessiert. Dieser maskiert jene Sequenzbereiche der Enzymsequenzen, die eine ungewöhnliche Aminosäurezusammensetzung aufweisen und infolgedessen zu Bildung zufälliger Sequenzpaarungen tendiert, so daß diese beim Alignment unberücksichtigt bleiben. Im Anschluß daran werden die Ergebnisse beider „alle vs. alle"-Alignments miteinander abgeglichen.

Die Maskierung von Sequenzbereichen geringer kompositorischer Komplexität verhindert das Zustandekommen von Alignments, die auf zufälliger Sequenzähnlichkeit beruhen, kann aber auch das Alignmentergebnis tatsächlich homologer Sequenzen beeinflussen. Da jedoch im folgenden anhand der Position und Ausdehnung der lokalen Alignments auf die modulare Struktur der Enzymsequenzen geschlossen werden soll, ist dieser Effekt unerwünscht. Daher werden anhand des gefilterten Alignmentlaufes zunächst die Sequenzpaarungen ermittelt, deren Sequenzübereinstimmungen nicht allein auf zufälliger Sequenzähnlichkeit beruhen. Anschließend werden genau diese Sequenzpaarungen aus dem ungefilterten Alignmentlauf ausgelesen. Es werden also nur jene Sequenzpaarungen des ungefilterten Laufes verwendet, die auch bei

Verwendung des Sequenzfilters zustande kommen. Die so erhaltenen lokalen Sequenzalignments dienen schließlich zur Ermittlung der Domänenstruktur aller im Datensatz enthaltenen Enzymsequenzen, während die beim Alignment ermittelten Werte für die statistische Signifikanz der Sequenzidentitäten zur anschließenden Gruppierung (Clusterung) der detektierten Domänen verwendet werden.

Anhand des Ergebnisses der Clusterung wird schließlich eine Auswahl von Clustern getroffen, deren Enzymzusammensetzung aufgrund ihrer funktionellen Diversität interessant erscheint und daher näher untersucht werden soll.

2.2 Implementierung und Speicherung

Das in dieser Arbeit entwickelte Programm wurde in der Programmiersprache *Python* implementiert. Eine Auflistung der verwendeten Module und benötigten Bibliotheken finden sich im Anhang A dieser Arbeit. Bei Verwendung bestehender Programme, die nicht auf eigenen Entwicklungen beruhen, wird explizit darauf verwiesen.

Die lokale Speicherung sämtlicher während eines Programmlaufs extrahierten und generierten Daten erfolgt in dem relationalen Datenbanksystem *MySQL* (Version 3.23.45). Dies erleichtert den Zugriff auf die umfangreiche Datensammlung und ermöglicht überdies eine Verknüpfung der unterschiedlichen Informationen. So können Zusammenhänge zwischen den unterschiedlichen Enzymklassen leichter aufgedeckt werden. Ein Datenbankschema findet sich im Anhang unter B.

2.3 Erstellung eines Enzymdatensatzes

2.3.1 Verwendete Sequenzdatenbanken

Zur Erstellung eines Datensatzes klassifizierter Enzyme werden die Proteinsequenzdatenbanken Swiss-Prot und TrEMBL [37] verwendet.

Swiss-Prot ist eine kurierte Sequenzdatenbank mit geringer Redundanz und umfangreicher, manuell erstellter Annotation. So enthält sie neben der reinen Sequenzinformation beispielsweise auch Angaben zur biologischen Quelle (Or-

ganismus, Gewebe), der Funktion, zu eventuell benötigten Cofaktoren oder posttranslationaler Modifikationen. Darüber hinaus umfaßt sie eine Vielzahl von Verknüpfungen zu anderen Datenbanken.

Die Proteinsequenzdatenbank TrEMBL (Translated EMBL) ist eine Computer-annotierte Ergänzung zu Swiss-Prot und enthält all jene in den Proteincode übersetzten Nukleotidsequenzen aus EMBL, die noch nicht Bestandteil von Swiss-Prot sind.

2.3.2 Aktualisierung der Datenbankversionen

Das implementierte Programm ist so konzipiert, daß es - als *cronjob*[2] gestartet - in regelmäßigen Intervallen den FTP-Server von EXPASY (Expert Protein Analysis System) [39] kontaktiert und prüft, ob neuere als die zuletzt verwendeten Versionen der Sequenzdatenbanken verfügbar sind. Erfahrungsgemäß ist dies etwa alle 12 bis 14 Tage der Fall.

Der Vergleich der Datenbankversionen erfolgt über die Dateigrößen. Die lokal abgelegte Datei „.former_sizes.tmp“ enthält die Größen der zuletzt verarbeiteten Datenbankversionen. Diese werden mit den Dateigrößen der auf dem Server verfügbaren Datenbanken verglichen. Unterscheiden sich die Werte, werden die aktuellen Fassungen automatisch zur Verarbeitung heruntergeladen und die lokale Referenzdatei wird entsprechend der neuen Dateigrößen aktualisiert. Selbiges geschieht auch, wenn keine lokale Referenzdatei zum Vergleich vorliegt, wie dies zum Beispiel beim ersten Programmlauf der Fall ist.

Neben den beiden Sequenzdatenbanken wird überdies die Datei „reldate.txt“ vom Server heruntergeladen, die Informationen zur jeweiligen Versionsnummer der Datenbanken enthält. Alle in dieser Arbeit präsentierten Ergebnisse beruhen auf Swiss-Prot Version 43.0 und TrEMBL in Version 25.13.

2.3.3 Extraktion klassifizierter Enzymsequenzen

Swiss-Prot und TrEMBL enthalten eine Vielzahl unterschiedlichster Proteine. Um unter diesen die Enzymsequenzen zu identifizieren, werden sukzessiv alle

[2] Der *Cron Daemon* ist eine Jobsteuerung von Unix bzw. Unix-artigen Betriebssystemen wie Linux, die wiederkehrende Aufgaben oder Befehle zu einer bestimmten Zeit ausführen kann.

Sequenzeinträge beider Datenbanken mittels eines *regulären Ausdrucks*[3] auf das Vorkommen einer oder mehrerer EC-Nummern geprüft. Verläuft die Prüfung erfolgreich, kommen weitere Kriterien zur Anwendung.

Da möglichst nur solche Enzymsequenzen Bestandteil des Datensatzes werden sollen, deren Klassifizierung experimentell ermittelt wurde, bleiben Enzymeinträge, deren deskriptives Feld (DE) darauf schließen läßt, daß es sich lediglich um eine hypothetische Zuordnung handelt, unberücksichtigt. Des weiteren gilt ein Längenkriterium, das Sequenzfragmente ausschließt, bzw. eine Mindestlänge garantiert. Tabelle 2-1 enthält eine Zusammenfassung aller für die Erstellung des Datensatzes geltenden Kriterien.

Tabelle 2-1: Kriterien für die Erstellung des Enzymdatensatzes

EC-Klassifikation	• mindestens eine vollständige, vierstellige EC-Nummer ohne unspezifizierte Ziffern (-) • die Beschreibung des Sequenzeintrages enthält keinen der Begriffe HYPOTHETICAL, POSSIBLE, POTENTIAL, PROBABLE oder PUTATIVE
Länge	• die Beschreibung des Sequenzeintrages enthält nicht den Begriff FRAGMENT • die Sequenzlänge beträgt mindestens 100 Aminosäuren

2.3.4 Aktualisierung der EC-Klassifikation

Die EC-Klassifikation unterliegt ständigen Änderungen und Aktualisierungen. Um sicherzustellen, daß die verwendete Einteilung tatsächlich der derzeit aktuellen entspricht, werden alle im Datensatz vorkommenden EC-Nummern auf ihre Gültigkeit geprüft und gegebenenfalls aktualisiert.

Die Aktualisierung erfolgt mittels einer manuell erstellten Referenzdatei (transferred_ec_numbers.dat). Die IUBMB veröffentlicht in unregelmäßigen Abständen Ergänzungen (*supplements*) zur bestehenden Klassifikation. Da diese aber weder einer vorgegebenen Form noch einem kontrollierten Vokabular unterliegen, müssen Änderungen zur Zeit manuell ergänzt werden. Der derzeitige

[3] Die Verwendung *Regulärer Ausdrücke* ermöglicht es dem Anwender, gezielt nach Zeichenketten zu suchen, die einem bestimmten Muster entsprechen.

Stand der zur Aktualisierung verwendeten Datei entspricht *Supplement 1* bis *9* der IUBMB.

Sequenzen, die lediglich von einer EC-Nummer zu einer anderen transferiert wurden, werden entsprechend aktualisiert, wobei auch mehrfach aufeinanderfolgende Änderungen berücksichtigt werden. Wurden hingegen bisher unter einer gemeinsamen EC-Nummer geführte Sequenzen auf mehrere unterschiedliche EC-Nummern verteilt, so kann zu diesem Zeitpunkt nicht entschieden werden, welcher neuen EC-Nummer eine jeweilige Sequenz nun angehört. Daher wird der bisherigen Nummer lediglich eine führende Null zur Kennzeichnung vorangestellt (z. B. 02.1.1.24). Die endgültige Zuordnung der Sequenzen ergibt sich in diesem Fall bei der abschließenden Clusterung. Sequenzeinträge, deren EC-Nummer ersatzlos gelöscht wurde, werden aus dem Datensatz entfernt.

2.4 Sequenzalignment

2.4.1 Theoretische Grundlagen

Das Sequenzalignment dient als Methode, um homologe Proteinsequenzen zu identifizieren. Dazu wird die zu betrachtende Sequenz (*Query*) mit allen Sequenzen einer spezifizierten Datenbank verglichen. Die Sequenzen werden einander gegenübergestellt und so gegeneinander ausgerichtet, daß möglichst viele ähnliche Aminosäurepositionen miteinander gepaart (aligniert) werden. Die Ähnlichkeit zwischen zwei Aminosäuren bezieht sich hierbei auf ihre physikalisch-chemischen Eigenschaften. Zu deren Bewertung werden sogenannte Substitutionsmatrizen verwendet, die die Ähnlichkeit der einzelnen Aminosäuren zueinander beschreiben. Anhand dieser Matrizen wird jeder möglichen Aminosäurepaarung ein festgelegter Ähnlichkeitswert (*score*) zugewiesen.

Positionen, die aufgrund von Insertionen bzw. Deletionen in einer der beiden Sequenzen nicht mit einer entsprechenden Position in der andern Sequenz gepaart werden können, werden als *gaps* bezeichnet. Um die Anzahl sowie die Länge der *gaps* in einem biologisch vertretbaren Rahmen zu halten, werden diese mit „Strafpunkten“ belegt. Da ein einzelnes Mutationsereignis die Insertion bzw. Deletion von mehr als einem Aminosäurerest verursachen kann, wird der Existenz einer Insertion/Deletion mehr Bedeutung zugeschrieben als deren

Länge. Daher wird das Einfügen eines *gaps* mit einem hohen Strafwert belegt, während die Verlängerung um jeden weiteren Aminosäurerest einen geringeren Strafwert erhält. Die Ähnlichkeit zweier Sequenzen ergibt sich schließlich als Summe der *scores* der gepaarten Aminosäuren abzüglich der Insertions- und Deletionsstrafen.

Prinzipiell können zwei Klassen von Alignmentalgorithmen unterschieden werden: lokale und globale. Im ersten Fall werden optimal passende Subsequenzen ermittelt, im letzteren die Ähnlichkeit über die gesamte Sequenz. Da viele Enzyme eine modulare Struktur und infolgedessen nur partielle Ähnlichkeit besitzen, wird in der vorliegenden Arbeit ein lokaler Alignmentalgorithmus verwendet.

2.4.2 Basic Local Alignment Search Tool

Zur Durchführung der Sequenzvergleiche wird eine lokale Installation des *Basic Local Alignment Search Tools*, kurz BLAST, in der Version 2.2.6 verwendet, das am *National Center for Biotechnology Information* (NCBI) entwickelt wurde [40].

Das BLAST-Softwarepaket umfaßt eine Reihe verschiedener Programme, die für den Vergleich unterschiedlicher Sequenzarten ausgelegt sind. Im vorliegenden Fall wird das Programm BLASTP verwendet, das für den Vergleich einer Proteinsequenz gegen eine Proteinsequenzdatenbank konzipiert ist.

Im Gegensatz zur Internetversion von BLAST, bei der nur vorgegebene Sequenzdatenbanken nach homologen Sequenzen durchsucht werden können, ermöglicht es eine lokale Installation, eigene Sequenzdatenbanken zu erstellen, die ausschließlich vom Anwender spezifizierte Sequenzen enthalten.

Da die von BLASTP erzeugten lokalen Alignments die Grundlage für alle weiteren Schritte der Methode bilden - von der Ermittlung der Domänenstruktur bis hin zur Clusterung der korrespondierender Sequenzabschnitte – soll im folgenden die Funktionsweise von BLASTP kurz erläutert werden.

2.4.2.1 Der BLASTP-Algorithmus

BLASTP basiert auf einem heuristischen Verfahren und verwendet eine „Wort"-basierende Suchprozedur, um Bereiche lokaler Sequenzähnlichkeit zu identifizieren. Der Algorithmus gliedert sich in vier Schritte:

1) Verarbeitung der Suchsequenz

Um die Sequenz eines Proteins gegen die Einträge einer Sequenzdatenbank zu vergleichen, ermittelt BLASTP zunächst eine Liste aller möglichen Teilsequenzen (vom BLAST-Autor als „Worte" bezeichnet) der Länge w, die in der Suchsequenz (*Query*) enthalten sind (vgl. Abbildung 2-2). Für Proteinsequenzen beträgt die Länge w 3 Aminosäuren [40]. Die maximale Anzahl möglicher Worte ergibt sich aus L-w+1, wobei L der Länge der Suchsequenz entspricht.

```
Query:      GSVEDTTGSQSLAANKCKTPQGQRRLVNQWIKQPLMDKNRI...
            GSV
             SVE
              VED
               EDT
                etc...
```

Abbildung 2-2: BLASTP erstellt zu Beginn eine Liste aller in der *Query* enthaltenen Teilsequenzen der Länge w = 3

Parallel dazu werden alle Aminosäurekombinationen der Länge w = 3 generiert, die mittels der 20 proteinogenen Aminosäuren gebildet werden können (20^3 = 8000 Kombinationen) und unter Verwendung eines Grenzwertes T mit jedem Wort der Suchsequenz verglichen. Für jede Position der Suchsequenz ergibt sich so eine Liste von Worten, die, wenn sie anhand der verwendeten Ähnlichkeitsmatrix (standardmäßig die Blocks Substitution Matrix 62, kurz BLOSUM62 [41]) mit dem Wort der Suchsequenz verglichen werden, mindestens einen *score* von T = 11 erreichen (vgl. Abbildung 2-3) [40]. Die ähnlichen Worte werden auch als „Nachbarn" bezeichnet.

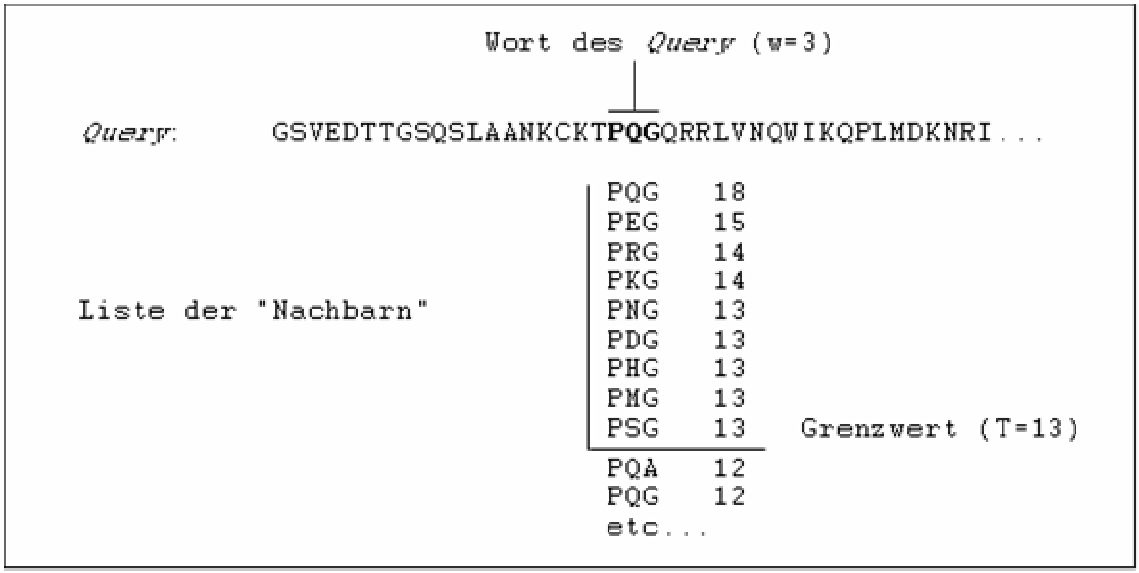


Abbildung 2-3: BLASTP erstellt für jedes Wort des *Query* eine Liste von „Nachbarn" der Länge w = 3, die mindestens einen Ähnlichkeitswert von T erreichen, wenn sie mit dem Wort des *Query* verglichen und mittels der Ähnlichkeitsmatrix bewertet werden. Dieser Prozeß ist zur Anschauung für das Wort „PQG" und einem Grenzwert von T = 13 dargestellt. Die zur Bewertung verwendete Ähnlichkeitsmatrix ist die Blocks Substitution Matrix 62, kurz BLOSUM62 [41].

2) Lokalisation identischer Worte in den Datenbanksequenzen

Nach dem ersten Schritt wird die Suchsequenz nun durch die Liste der „Nachbarn" für jede Position repräsentiert. Um die Suchsequenz mit einer Sequenz der Datenbank zu vergleichen, wird nach Übereinstimmungen zwischen den „Nachbarn" und den Worten der Datenbanksequenz gesucht. Wenn ein Wort der „Nachbarn" identisch mit einem Wort der Datenbanksequenz ist, wird ein Treffer (*hit*) verzeichnet. Auf diese Weise werden alle *hits* zwischen Suchsequenz und Datenbanksequenz ermittelt. Ein *hit* ist jeweils durch seine Position in jeder der beiden Sequenzen charakterisiert.

3) Auswahl und Erweiterung primärer *hits*

Um festzustellen, ob ein *hit* Teil eines größeren Bereichs der Sequenzübereinstimmung zwischen Suchsequenz und der Datenbanksequenz ist, wird dieses Segment kurzer Übereinstimmung ohne Einfügen von *gaps* in beide Richtungen verlängert. Die Verlängerung erfolgt solange, bis der für die Sequenzübereinstimmung errechnete Ähnlichkeitswert – in Bezug auf den höchsten Wert, der

bis dahin während der Verlängerung erreicht wurde – um mehr als *X* abfällt. Der Wert *X* ist ein Parameter des Programms und kann vom Anwender bestimmt werden. Das Resultat der Verlängerung ist ein lokaler Bereich optimaler Sequenzpaarung, der mindestens einen *hit* beinhaltet. Liegt der dafür errechnete *score* über einem gesetzten Grenzwert, wird er als *High-scoring Segment-Pair* (*HSP*) bezeichnet.

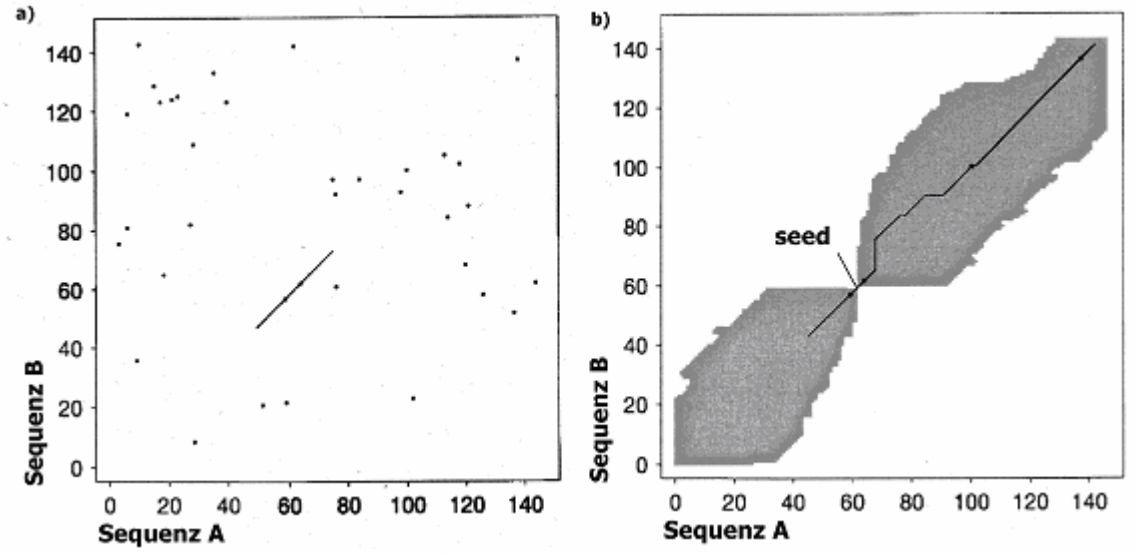


Abbildung 2-4: a) BLAST identifiziert zunächst primäre *hits* (durch Punkte dargestellt). Im vorliegenden Beispiel ist lediglich ein Paar von *hits* auf derselben Diagonale lokalisiert und besitzt einen Abstand kleiner *A*. Dieses wird zu einem *HSP* (*High-scoring Segment-Pair*) verlängert. b) Für *HSP*s mit einem *score* größer S_g wird ein *gapped*-Alignment berechnet. Die Berechnung beginnt vom Aminosäurepaar, das als *seed* bezeichnet wird (siehe unten) und endet, wenn der *score* des Alignment unter einen vorgegebenen Grenzwert abfällt.

Hits, die nicht durch eine Verlängerung ohne *gaps* miteinander verbunden werden können, können Teil desselben mit *gaps* versehenen Alignments sein. Daher müssen nicht notwendigerweise alle zuvor ermittelten primären *hits* verlängert werden. Um die Geschwindigkeit von BLASTP zu erhöhen, erfolgt daher eine Auswahl von primären *hits*, die für eine Verlängerung in Frage kommen. Für die Auswahl gelten folgenden Kriterien: Es muß zu einem *hit* einen weiteren nicht-überlappenden *hit* geben, der - in Bezug auf die Matrix des Smith-Waterman-Algorithmus [11] - auf derselben Diagonalen lokalisiert ist (vgl. Abbildung 2-4a). Der räumliche Abstand dieser *hits* zueinander muß kleiner *A* sein, wobei *A* den Abstand zwischen den ersten Positionen der beiden *hits* be-

zeichnet. Für den Vergleich von Proteinsequenzen beträgt *A* standardmäßig 40 [40]. Alle *hits*, die diese Kriterien erfüllen, werden wie beschrieben zu einem *HSP* verlängert.

4) Verknüpfung der *HSP*s unter Berücksichtigung möglicher Insertionen und Deletionen

Die aus den *hits* generierten *HSP*s dienen, sofern sie einen vorgegebenen Grenzwert S_g erreichen, als Ausgangspunkt für die Ausführung eines dynamisch programmierten lokalen Alignments. Der Algorithmus, der zur Berechnung dieser lokalen, mit *gaps* versehenen Alignments verwendet wird, ist eine Modifikation des Smith-Waterman-Algorithmus [11].

Anfangs wird ein Startpunkt (*seed*) innerhalb des *HSP*s festgelegt (vgl. Abbildung 2-4b). Dazu wird innerhalb des *HSP*s das Segment der Länge 11 bestimmt, daß den höchsten *score* erreicht [40]. Das mittlere Aminosäurepaar dieses Elfer-Segments ist der Startpunkt. Wenn das *HSP* kürzer als 11 sein sollte, wird dessen zentrales Aminosäurepaar als *seed* gewählt. Von diesem Punkt ausgehend wird das Alignment in beide Richtungen verlängert. Die Suche nach dem optimalen Verlauf ist auf bestimmte Bereiche der Matrix beschränkt, so daß der *score* des Alignments - im Vergleich zum besten Wert, der bis dahin während der Verlängerung erreicht wurde - nicht um mehr als den Wert X_g abfällt [40].

2.4.2.2 Bewertung der Alignments

BLASTP errechnet für jeden Sequenzvergleich vier Werte, die den Grad der Ähnlichkeit zwischen Suchsequenz und Datenbanksequenz widerspiegeln:

- *identity*: Die Identität gibt den prozentualen Anteil identischer Aminosäurepositionen zwischen *Query* und Datenbanksequenz im alignierten Sequenzbereich an.
- *raw score*: Der *raw score* eines Alignments errechnet sich als die Summe der *scores* der alignierten Aminosäurepositionen abzüglich der Insertions- und Deletionsstrafen. Je höher der Alignment-*score*, desto größer ist die Sequenzähnlichkeit. Der *score* s_{ij} eines alignierten Aminosäurepaares *i* und *j*

wird der verwendeten Substitutionsmatrix entnommen, in der jeder möglichen Aminosäure-Substitution unter Berücksichtigung ihrer physikalisch-chemischen Eigenschaften ein Ähnlichkeitswert zugewiesen wird. BLASTP verwendet standardmäßig die Blocks Substitution Matrix 62, kurz BLOSUM62 [41].

- *bit score*: Das zugrundeliegende statistische Modell basiert auf der Annahme, daß jede Aminosäure *i* in allen Positionen der Proteinsequenz mit der gleichen Wahrscheinlichkeit P_i auftritt. Überdies wird angenommen, daß der erwartete *score* für zwei zufällige Aminosäuren (vgl. Gleichung **1**) negativ ist.

$$\sum_{ij} P_i P_j s_{ij} \qquad \textbf{(1)}$$

Bei gegebenen Werten für P_i und s_{ij}, bietet die zugrundeliegende Theorie [15] zwei Parameter λ und *K*, die genutzt werden können, um den *raw score* in einen normalisierten *score* umzuwandeln. Normalisierte *scores* können im Gegensatz zu *raw scores* unabhängig vom verwendeten Bewertungssystems direkt miteinander vergleichen werden. Der normalisierte *score* S_n ergibt sich aus:

$$S_n = \frac{\lambda S - \ln K}{\ln 2} \qquad \textbf{(2)}$$

Wenn der Logarithmus zur Basis 2 zur Berechnung verwendet wird, ist die Einheit des scores *bit* [15, 42].

- *E-value:* Um signifikante Sequenzidentität von zufälliger Ähnlichkeit zu unterscheiden, wird die statistische Signifikanz jedes Alignments in Form eines Erwartungswertes (*Expectation value* bzw. *E-value*) berechnet. Der *E-value* für ein Alignment mit dem *score S* gibt an, wie häufig Alignments mit gleichem oder einem besseren *score* als *S* zu erwarten sind, wenn zufällig gewählte Sequenzen der gleichen Länge und Aminosäurezusammensetzung wie Suchsequenz und Datenbanksequenz miteinander verglichen werden. Je niedriger der *E-value*, desto signifikanter ist das ermittelte

Alignment. Wenn zwei zufällige Sequenzen der Längen *m* und *n* miteinander verglichen werden, ergibt sich die Anzahl *E* der *HSP*s mit einem normalisierten *score* von mindestens S_n, deren Auftreten zufällig erwartet wird, aus der Näherungsgleichung:

$$E = \frac{N}{2^{S_n}} \qquad \textbf{(3)}$$

N entspricht hierbei dem Produkt aus *m* und *n*. Gleichung **3** kann nach $S_n = log_2 (N / E)$ umgeformt werden, woraus sich der normalisierte *score* S_n ergibt, der einen bestimmten *E-value* entspricht.

2.4.2.3 Ausführung der Sequenzalignments

Während eines Programmlaufes werden zwei vollständige „alle vs. alle"-Alignments ausgeführt, bei denen jeweils jede Sequenz paarweise mit allen anderen des Datensatzes verglichen wird. Die Ausführung erfolgt mittels zweier Shell-Skripte und unter Verwendung unterschiedlicher Parameter.

Für eines der Alignments werden die Sequenzen zuvor mit dem in BLASTP integrierten Sequenzfilter SEG [43, 44] prozessiert. Dieser Filter dient zur Maskierung von Sequenzbereichen geringer kompositorischer Komplexität (low complexity region) und verhindert so, daß diese in das Alignment eingehen. Das andere Alignment erfolgt unter Verwendung der unveränderten Originalsequenzen, wie sie aus Swiss-Prot und TrEMBL extrahiert wurden.

Um die für die Alignments benötigte Rechenzeit zu reduzieren und innerhalb eines „alle vs. alle"-Alignments wiederholte Vergleiche derselben Sequenzpaarung zu vermeiden, werden die Sequenzen nach Länge getrennt prozessiert, so daß die Sequenzen einer jeweiligen Länge immer nur gegen solche gleicher oder geringerer Sequenzlänge verglichen werden. Um trotz wechselnder Größe der zum Vergleich verwendeten Sequenzdatenbank vergleichbare *E-values* zu erhalten (vgl. Abschnitt 2.4.2.2), wird die tatsächliche Gesamtgröße des Sequenzdatensatzes als Parameter an BLAST übergeben und zur Berechnung des *E-values* verwendet.

Eine weitere Reduktion der Rechenzeit wird dadurch erzielt, daß die Suchsequenzen einer jeweiligen Sequenzlänge auf Dateien mit je 25 Sequenzen verteilt werden, von denen mittels *Threading* je vier zeitgleich verarbeitet werden.

Nach Abschluß beider „alle vs. alle"-Sequenzalignments werden anhand des mit SEG gefilterten Alignments zunächst die Sequenzpaarungen ermittelt, deren Sequenzähnlichkeit nicht ausschließlich auf Sequenzbereiche geringer kompositorischer Komplexität beruhen. Anschließend werden die Sequenzpaarungen des ungefilterten Alignments ausgelesen und mit denen des gefilterten Alignments abgeglichen. Um für eine weiter Verarbeitung in Frage zu kommen, muß eine Sequenzpaarung des ungefilterten Alignmentlaufes folgende Kriterien erfüllen:

- die Paarung kommt auch bei Verwendung des Sequenzfilters zustande
- der für die Sequenzpaarung errechnete *E-value* ist kleiner als 10^{-1}
- die miteinander gepaarten Sequenzbereiche beider Sequenzen umfassen mindestens 60 Aminosäuren

2.5 Ermittlung der Domänenstruktur

Die Bestimmung der Sequenzdomänenstruktur erfolgt mittels der Position und Ausdehnung der von BLASTP erzeugten lokalen Alignments und verläuft in drei Schritten:

1) Zunächst werden alle zu einer Sequenz verfügbaren Alignments ermittelt. Dies schließt auch jene Alignments ein, bei denen die Sequenz nicht als Such- (*Query*), sondern als Datenbanksequenz fungiert. Anschließend werden die N- und C-Termini der lokalen Alignments auf die Länge der zu betrachtende Sequenz projiziert (vgl. Abbildung 2-5). Infolgedessen ergibt sich ein charakteristisches Peak-Muster. Die Positionen der Peaks kennzeichnen die Lage der N- und C-Termini der lokalen Alignments innerhalb der Sequenz, während ihre Höhe die Häufigkeit angibt, mit der eine Position als Anfang bzw. Ende eines Alignments dient. Ist eine

Sequenzposition sowohl Start- als auch End-Terminus von Alignments, so ergibt sich die Höhe ihres Peaks als Summe beider Häufigkeiten.

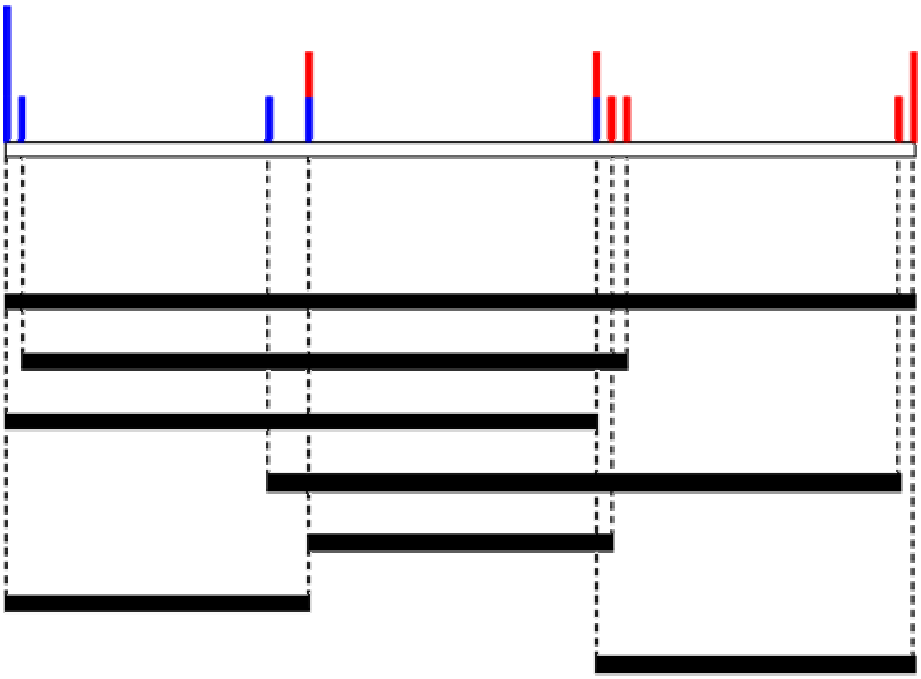


Abbildung 2-5: Die graphische Darstellung zeigt anhand eines fiktiven Beispiels, wie die Start- (blau) und End-Termini (rot) der zu einer Sequenz verfügbaren lokalen Alignments (■) auf die betrachtete Sequenz (❑) übertragen werden.

Die nun folgenden zwei Schritte dienen dazu, das Muster der Peaks zu „glätten“, um so die Grenzen zwischen den Domänen herauszuarbeiten. Zu diesem Zweck werden alle Alignmentregionen, die im wesentlichen denselben Sequenzabschnitt umfassen, zu einer einzigen Region zusammengefaßt, die fortan all diese Varianten repräsentiert. Infolgedessen verringert sich die Gesamtzahl der zu betrachtenden Regionen.

2) Während des ersten Schrittes der Zusammenfassung werden N- und C-Termini der Alignments getrennt voneinander betrachtet. Dazu wird ein Raster mit einer Fenstergröße von 15 Aminosäuren über die Sequenz gelegt. Beginnend mit den N-Termini der Alignments werden alle Peaks, die innerhalb desselben „Fensters“ liegen, zu einem einzelnen Peak vereint, der von da an alle N-Termini innerhalb dieses Bereiches repräsentiert. Die Position des neuen Peaks entspricht der Sequenzposition innerhalb des „Fensters“, die am häufigsten als N-Termini dient. Die

Höhe des neuen Peaks ergibt sich indes als die Summe aller Häufigkeiten der vereinten Peaks (vgl. Abbildung 2-6a). Mit den C-Termini der Alignments wird anschließend auf dieselbe Weise verfahren. Existieren innerhalb eines „Fensters“ mehrere Maxima gleicher Größe, so werden die daran angrenzenden Positionen zur Bestimmung der Lage des neuen Peaks herangezogen (vgl. Abbildung 2-6b). Zunächst werden die Häufigkeitswerte der jeweils direkt links und rechts an die Maxima angrenzenden Positionen ermittelt und deren jeweilige Summen miteinander verglichen. Sollte auch dies kein Entscheidung bringen, wird der betrachtete Bereich sukzessive um je eine weitere Position in beide Richtungen erweitert, bis schließlich eindeutiges Maximum bestimmt oder aber das Ende der Sequenz erreicht ist. In letzteren Fall wird automatisch die Position des Maximums gewählt, die dem Sequenzende am nächsten liegt.

3) Im letzten Schritt des Verfahrens werden die Alignmentregionen nicht mehr nach N- und C-Termini getrennt verarbeitet, sondern in ihrer Gesamtheit betrachtet. Nachdem im vorherigen Schritt alle geringfügigen Variationen der Start- bzw. End-Termini durch einzelne Repräsentanten ersetzt wurden, werden nun größere Variationen der Domänenlänge berücksichtigt. Ziel ist es, alle Längenvarianten einer Domäne zusammenzufassen, ohne dabei benachbarte Domänen zu inkorporieren. Zu diesem Zweck werden alle deckungsähnlichen Alignmentregionen unter Verwendung der Clusteranalyse gruppiert. Eine detaillierte Beschreibung der Vorgehensweise findet sich im Abschnitt „Clusteranalyse“ im Abschnitt 2.6.2.1.

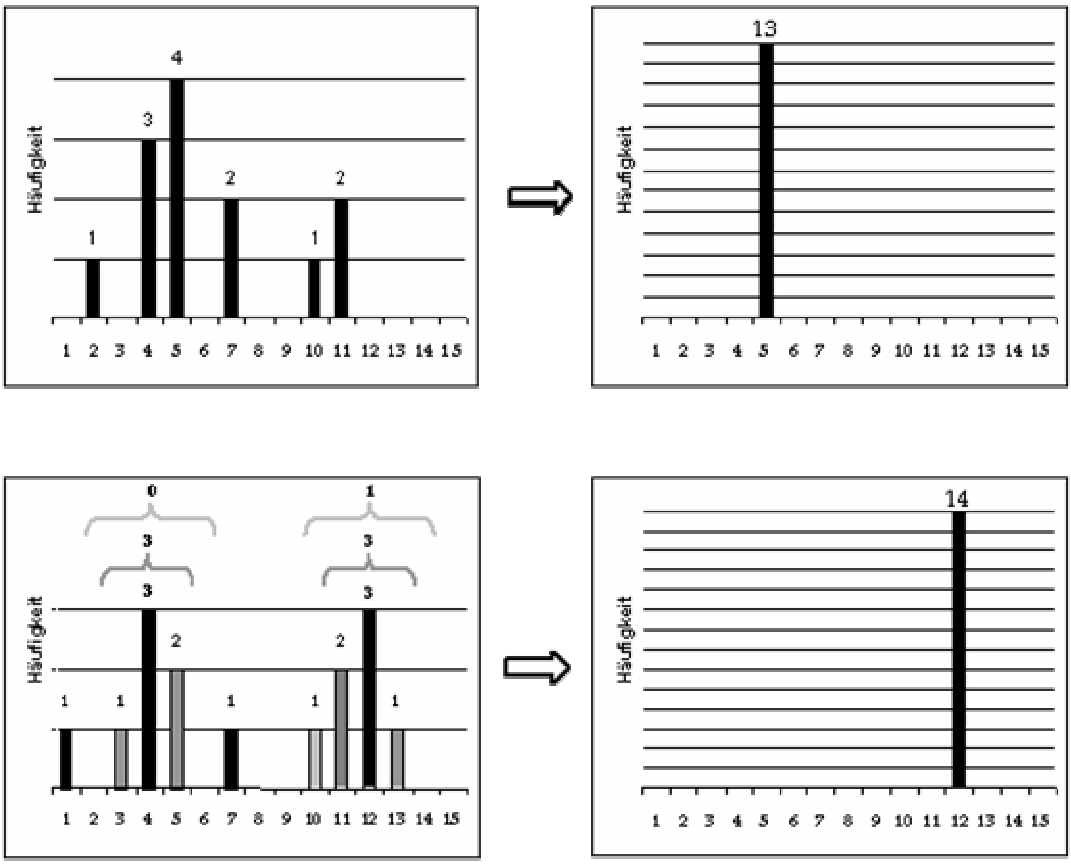


Abbildung 2-6: a) Es wird ein Raster mit einer Fenstergröße von 15 Aminosäuren über die zu betrachtende Sequenz gelegt. Beginnend mit den N-Termini der Alignments werden alle Peaks, die innerhalb desselben „Fensters“ liegen, zu einem einzelnen Peak vereint, der von da an alle N-Termini innerhalb dieses Bereiches repräsentiert. Die Position des repräsentierenden Peaks entspricht der Position mit der größten Häufigkeit innerhalb des „Fensters“, während sich seine Größe als Summe der Häufigkeiten aller vereinten Peaks ergibt. Mit den C-Termini der Alignments wird anschließend auf dieselbe Weise verfahren. b) Bei mehreren identischen Maxima innerhalb eines „Fensters“ werden die flankierenden Positionen zur Entscheidungsfindung hinzugezogen. Dazu werden ihre Häufigkeitswerte ermittelt und die Summen miteinander verglichen. Führt auch dies nicht zu einer Entscheidung, wird der betrachtet Bereich sukzessive um je eine weitere Positionen in beide Richtungen erweitert (durch Klammern angedeutet), bis ein eindeutiges Ergebnis erzielt wird.

2.6 Clusteranalyse

2.6.1 Theoretische Grundlagen

Die Literatur unterscheidet zwischen hierarchischen und nicht hierarchischen Clusterverfahren, wobei erstere nochmals nach divisiven und agglomerativen unterschieden werden. Da in der vorliegenden Arbeit nur das hierarchisch agglomerative Verfahren verwendet wurde, soll hier nur auf dieses eingegangen werden.

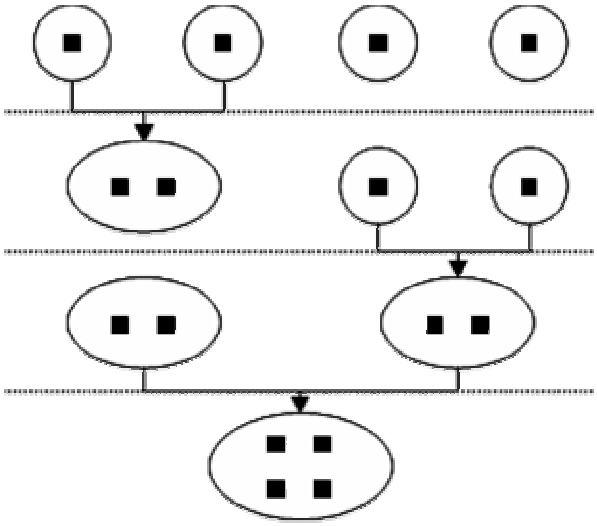


Abbildung 2-7: Graphische Repräsentation des Gruppierungsverlaufs beim agglomerativ hierarchischen Clusterverfahren (○: Cluster; ■: zu gruppierende Objekte). Der Prozeß kann solange fortgeführt werden, bis ein vorgegebenes Abbruchkriterium erfüllt ist (gestrichelte Linien), oder alle Objekte in einem einzigen Cluster enthalten sind.

Beim hierarchischen Klassifizierungsverfahren wird eine baumähnliche, hierarchische Struktur der Objekte erstellt. Bei der Clusterung wird iterativ vorgegangen, d. h. die Gruppierung entsteht schrittweise durch Vergleich jedes einzelnen Objekts mit allen anderen Objekten bzw. mit den in den vorangegangenen Teilschritten bereits aufgebauten Gruppierungen.

Die Ausgangsposition ist dadurch charakterisiert, daß jedes einzelne Objekt zunächst als ein eigenständiges Cluster (Einheitscluster) verstanden wird, das im zweiten Schritt mit dem ihm ähnlichsten Einheitscluster zu einem Zweielement-Cluster zusammengeführt wird (vgl. Abbildung 2-7). Im darauffolgenden Schritt wird diesem neuen Cluster ein Objekt hinzugefügt oder es werden zwei

weitere Einheitscluster miteinander fusioniert. Durch wiederholtes Berechnen der Distanz und Fusionierung der „nächsten“ Objekte, respektive Cluster entsteht dabei ein binärer Baum, der die Distanz der einzelnen Objekte zueinander veranschaulicht.

Voraussetzung für eine Clusterung ist, daß man ein quantitatives Maß für den Grad der Ähnlichkeit zwischen den Objekten hat, anhand dessen eine Einteilung erfolgen kann. Darüber hinaus wird aber noch ein weiters Distanzmaß benötigt. Läßt sich der Abstand zweier Objekte eindeutig festlegen, so gilt dies für die Distanz zwischen Clustern, die mehrere Objekte enthalten, nicht. Vielmehr existieren eine Vielzahl denkbarer Möglichkeiten, den Abstand zwischen den einzelnen Gruppen von Objekten zu definieren. Die Wahl des Kriteriums bestimmt im wesentlichen die Art der Clusteranalyse, denn es entscheidet darüber, welche Cluster zusammengefügt werden. Je nachdem wie man den Abstand zweier Cluster definiert, können sich die Ergebnisse einer Clusteranalyse mitunter erheblich voneinander unterscheiden. Welches Distanzmaß am geeignetsten ist, ergibt sich aus den jeweiligen Anforderungen.

2.6.2 Ausführung der Clusterung

Die Methode der Clusteranalyse kommt im Verlauf des Programms zweimal zur Anwendung. Sie wird zum einen bei der Ermittlung der Sequenzdomänen und ihrer Grenzen und zum anderen bei der abschließenden Gruppierung korrespondierender Domänenabschnitte verwendet. Für die unterschiedlichen Aufgaben kommen verschiedene Distanzmaße zum Einsatz, um die Ähnlichkeit zweier Cluster zu beschreiben.

2.6.2.1 Ermittlung der Sequenzdomänen und ihrer Grenzen

Um die Sequenzdomänenstruktur der Enzyme zu ermitteln, wird das sogenannte *complete linkage* Verfahren verwendet. Für diese Methode ist die Distanz zwischen zwei Clustern als die längste Distanz zwischen einem Objekt im ersten und einem Objekt im zweiten Cluster definiert (vgl. Abbildung 2-8).

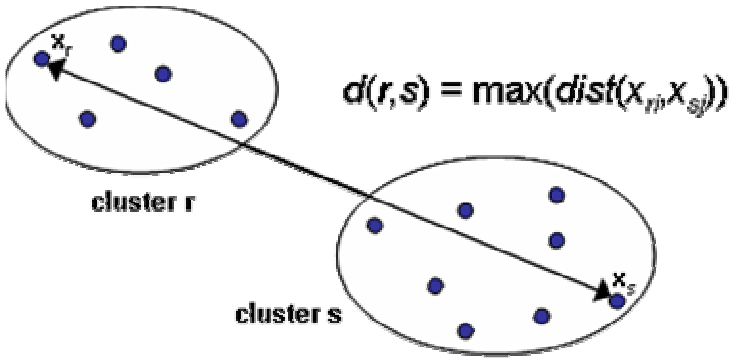


Abbildung 2-8: Graphische Repräsentation des Distanzmaßes zwischen zwei Clustern für die *complete linkage* Methode. Die Distanz der Cluster entspricht der längsten Distanz zwischen den Objekten beider Cluster.

Die Ausführung der Clusterung gestaltete sich wie folgt: Nachdem im ersten Schritt der Zusammenfassung alle geringfügigen Abweichung der Start- und End-Termini angeglichen wurden (vgl. Abschnitt 2.5), werden nun auch größere Variationen der Domänenlänge berücksichtigt. Dazu werden alle nach dem ersten Schritt verbliebenen Alignmentregionen zu einer Sequenz miteinander verglichen, um festzustellen, ob sie einander überlappen. Ist dies der Fall, so wird für diese die Anzahl der nicht-überlappenden Aminosäurereste bestimmt. Je geringer diese ausfällt, desto ähnlicher sind sich zwei Alignmentregionen. Anschließend werden die überlappenden Regionen schrittweise gruppiert, wobei mit den Alignmentregionen begonnen wird, die den geringsten Unterschied zueinander aufweisen. Um Regionen nicht schon bei geringfügiger Überlappung zu gruppieren, sondern den Prozeß auf deckungsähnliche Alignmentbereich zu beschränken, wird ein Schwellwert verwendet, der die maximal zulässige Abweichung definiert. Demnach werden nur solche Alignmentregionen in einem Cluster zusammengefaßt, die sich N- wie C-terminal jeweils nicht mehr als 45 Aminosäurereste voneinander unterscheiden.

Nachdem alle Regionen zu einer Sequenz auf diese Weise gruppiert worden sind, werden die in einem jeweiligen Cluster enthaltenen Varianten einer Alignmentregion durch eine einzige Region ersetzt, die fortan all diese Varianten repräsentiert (vgl. Abschnitt 2.5). Die Grenzen der neuen Region entsprechen dem N- bzw. C-Terminus mit der jeweils größten Häufigkeit innerhalb des Clusters. Anschließend wird die Prozedur für die so erhaltenen Regionen wiederholt. Dies erfolgt solange, bis schließlich keine Änderung mehr eintritt und

alle Längenvarianten einer Alignmentregion durch eine einzige Region repräsentiert werden.

2.6.2.2 Gruppierung korrespondierender Sequenzabschnitte homologer Enzyme

Die Clusterung der korrespondierenden Sequenzabschnitte in homologen Enzymen erfolgt unter Verwendung der *single linkage* Methode. Im Gegensatz zur *complete linkage* Methode definiert diese die Distanz zwischen zwei Clustern als die kürzeste Distanz zwischen einem Objekt im ersten und einem Objekt im zweiten Cluster (vgl. Abbildung 2-9).

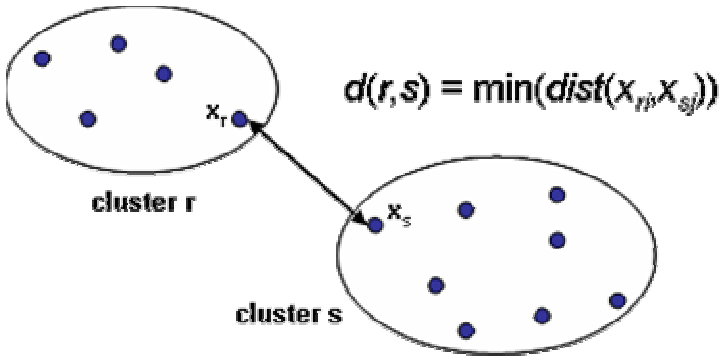


Abbildung 2-9: Graphische Repräsentation des Distanzmaßes zwischen zwei Clustern für die *single linkage* Methode. Die Distanz der Cluster entspricht der kürzesten Distanz zwischen den Objekten beider Cluster.

Nachdem bei der Ermittlung der Domänenstruktur alle deckungsähnlichen Alignmentregionen zu einer Sequenz durch einzelne Repräsentanten ersetzt worden sind, ergibt sich nun für jede Sequenz eine überschaubare Anzahl von verschiedenen Regionen, über die sie Ähnlichkeit zu anderen Sequenzen besitzt. Die Gesamtzahl all dieser Regionen für alle im Datensatz enthaltenen Sequenzen bildet die zu gruppierende Objektmenge.

Es wird mit der Sequenzpaarung begonnen, welche die größte Ähnlichkeit besitzt, wobei der beim Alignment ermittelte Wert für den *E-value* verwendet wird, um den Grad der Identität zu quantifizieren. Das dazugehörige lokale Alignment gibt Auskunft darüber, welche der für beide Sequenzen zur Verfügungen stehenden Regionen einander entsprechen. Diese werden daraufhin zu

einem Cluster zusammengefaßt. Im nächsten Schritt wird der Signifikanzlevel schrittweise erniedrigt, und es werden nach und nach weitere Ähnlichkeiten berücksichtigt.

Um die Entwicklung der Clusterbaumes verfolgen zu können, werden Zwischenergebnisse für insgesamt zehn verschiedene Grenzwerte gespeichert. Die Grenzwerte für den *E-value* betragen 10^{-180}, 10^{-110}, 10^{-80}, 10^{-50}, 10^{-35}, 10^{-25}, 10^{-15}, 10^{-10}, 10^{-5} und 10^{-2}.

2.7 Untersuchung der Clusterergebnisse

Anhand des Clusterergebnisses soll geklärt werden, ob eine auf Sequenzhomologie basierende Einteilung der Enzyme mit den von den Enzymen verwendeten Reaktionsmechanismen korreliert. Zur Klärung dieser Frage wird eine Auswahl von Clustern getroffen, deren Enzymzusammensetzung aufgrund ihrer funktionellen Diversität interessant erscheint und daher auf das Vorliegen ähnlicher Reaktionsmechanismen untersucht werden soll. Die Untersuchung der Cluster erfolgt anhand verschiedener Kriterien.

2.7.1 Identifizierung der Domänenfunktion

Zu Beginn der Untersuchung muß geklärt werden, ob die im Cluster enthaltenen Sequenzabschnitte tatsächlich die katalytischen Domänen der Enzyme umfassen. Dies geschieht mittels bestehender Sequenzdomänensammlungen.

Sequenzdomänen werden in Form ihrer Domänendeskriptoren von wenigen Institutionen katalogisiert und über das Internet zugänglich gemacht. Zu nennen ist das SIB (Swiss Institute of Bioinformatics), das die sogenannte PROSITE-Datenbank [45] pflegt und erweitert, und das Sanger-Zentrum, das mit der Pfam-Datenbank [46] eine noch größere Sammlung anbietet, die aber nicht immer die Sensitivität der analogen PROSITE-Profile erreicht. Beide Profilsammlungen überschneiden sich zum Teil, beinhalten aber auch für die jeweilige Sammlung spezifische Profile. Daneben existieren noch Profilsammlungen von SMART [47], PRINTS [48] und BLOCKS [49], die im Rahmen dieser Arbeit jedoch nicht verwendet wurden.

2.7.2 Identifizierung konservierter Sequenzpositionen

Wenn das zu betrachtende Cluster die katalytische Domänen der Enzyme enthält, wird ein multiples Sequenzalignment (MSA; eine Erweiterung des paarweisen Alignments, bei dem drei und mehr Sequenzen miteinander aligniert werden) der gruppierten Sequenzabschnitte erstellt.

Es stehen eine Reihe verschiedene Programme zur Erzeugung multipler Alignments zur Verfügung. Das verbreiteste unter ihnen ist ClustalW [50], das ein geschwindigkeitsoptimiertes, heuristisches Verfahren verwendet und daher auch für größere Sequenzzahlen einsetzbar ist. In der vorliegenden Arbeit wurde jedoch dem langsameren Programm TCoffee [51] der Vorzug gegeben, da dieses häufig Alignments besserer Qualität erzeugt [52]. Im Falle sehr großer Cluster, die mehrere hundert Sequenzen enthalten, wurde indes auf das Programm MUSCLE [53] zurückgegriffen, das sowohl in Geschwindigkeit als auch Alignmentqualität in etwa zwischen ClustalW und TCoffee rangiert.

Anhand der multiplen Sequenzalignments können konservierte oder konservativ substituierte Aminosäurepositionen identifiziert werden, die für die Funktion häufig von essentieller Bedeutung sind. So ergeben sich erste Hinweise auf möglicherweise katalytisch aktive Aminosäurereste.

2.7.3 Lage der konservierten Sequenzpositionen in der Proteinstruktur

Wenn für die im Cluster enthaltenen Sequenzen dreidimensionale Proteinstrukturen verfügbar sind, wird die Lage der im vorrangegangenen Schritt identifizierten Aminosäurepositionen innerhalb der Strukturen betrachtet.

Anhand der Strukturen kann überprüft werden, ob die konservierten Reste innerhalb der verschiedenen Enzyme in vergleichbarer Orientierung und Abstand zueinander vorliegen. Dies ist Voraussetzung für einen ähnlichen Reaktionsmechanismus.

3 Ergebnisse und Diskussion

3.1 Praktische und theoretische Probleme der Proteinklassifikation

Die klassische Methode, die Eigenschaften eines Proteins vorherzusagen, basiert auf seiner Ähnlichkeit zu den Sequenzen bereits charakterisierter Proteine [54]. Die zugrundeliegende Hypothese besagt, daß Eigenschaften wie Struktur und Funktion zwischen homologen Sequenzen transferiert werden können, da diese während der Evolution konserviert wurden. Diese Annahme kann im Fall der Proteinstruktur bestätigt werden, für die ein direkter Zusammenhang zwischen Sequenzähnlichkeit und Konservierung der Struktur nachgewiesen wurde [55, 56]. Im Fall der Proteinfunktion wird eine ähnliche Beziehung angenommen, erscheint aber weit weniger gerechtfertigt [57]. So wurde beispielsweise untersucht, inwieweit Sequenzähnlichkeit geeignet ist, um Enzymsequenzen einer spezifischen EC-Klasse zuzuordnen. Je nach verwendeten Methoden und Bewertungskriterien erwies sich Sequenzähnlichkeit jedoch in bis zu 60% der Fälle als unzureichend, um eine eindeutige Zuordnung vorzunehmen [58, 59].

Die Schwierigkeiten dieses Prozesses beruhen zum Teil auf der theoretischen Definition der Funktion, sind zum Teil aber auch auf praktische Probleme zurückzuführen [60]. Einige von ihnen, die sich aus den Analysewerkzeugen und Datenbanken ergeben, sind

- die Identifizierung ähnlicher Sequenzen in großen Datenbanken,
- systematische Fehler bei der Detektierung von Homologen,
- falsch annotierten Sequenzen in den Datenbanken und
- die modulare Domänenarchitektur vieler Enzyme.

Neben diesen praktischen Problemen erweist sich die Definition der Funktion selbst als problematisch. Die funktionelle Klassifizierung des EC-Systems ist eine empirische Einteilung und basiert auf den beobachteten Reaktionen. Sie ist somit unabhängig von der evolutionären Abstammung oder den Reaktionsmechanismen der Enzyme. Infolgedessen erhalten analoge Enzyme, welche die

gleiche Reaktion mittels verschiedener Mechanismen katalysieren, die gleiche EC-Klassifikation, während homologe Enzyme, die ähnliche Mechanismen zur Katalyse unterschiedlicher Reaktionen verwenden, verschiedenen EC-Klassen zugeordnet werden. Signifikante Sequenzähnlichkeit deutet also in erster Linie auf eine ähnliche Struktur, aber nicht notwendigerweise auf eine identische biochemische Funktion hin. Aufgrund von Sequenzhomologie läßt sich für ein Protein aber oftmals die Zugehörigkeit zu einer Proteinfamilie vorhersagen. Auch ist innerhalb einer Proteinfamilie häufig ein allgemeiner biochemischer Wirkmechanismus, wie zum Beispiel ein enzymatischer Reaktionsmechanismus oder die Rolle als Bindungspartner für einen Co-Faktor konserviert [61]. Diese Beziehung wird in der vorliegenden Arbeit genutzt, um – unter Vermeidung der zuvor genannten praktischen Probleme - eine auf Homologie basierende Einteilung der Enzyme zu erstellen, die der Hypothese nach auch mit den Reaktionsmechanismen der Enzyme korreliert.

3.2 Der Enzymdatensatz

Die im folgenden präsentierten Ergebnisse beruhen auf UniProt Release 1.6 bestehend aus Swiss-Prot in Version 43.0 und TrEMBL in Version 25.13 [37]. Diese Fassungen der Sequenzdatenbanken enthalten 146.720 bzw. 1.069.649 Sequenzeinträge. Anhand des *Keyword*-Feldes (KW) können insgesamt 144.299 Enzymsequenzen identifiziert werden.

3.2.1 Datensatz aus Enzymen mit experimentell bestimmter Funktion

Die gängige Praxis, Enzyme mit unbekannten Eigenschaften aufgrund von Sequenzähnlichkeit zu klassifizieren, hat dazu geführt, daß ein Großteil der zur Verfügung stehenden Annotationen hypothetischer Natur ist. Wird die Klassifizierung von einem versierten Anwender vorgenommen, ist diese Vorgehensweise meist unproblematisch. Kommen jedoch, wie bei der computergestützten Annotierung ganzer Genome, automatische Methoden zum Einsatz, so bleibt die Frage nach der Verläßlichkeit der erhaltenen Klassifikation.

Sequenzähnlichkeit ist, wie zuvor erwähnt, nur bedingt als Indikator für funktionelle Ähnlichkeit geeignet. Viele der automatisch generierten Annotationen können demzufolge falsch sein. Einmal etablierte Fehler können sich zu-

dem durch mehrmaliges Kopieren der Annotation zwischen ähnlichen Sequenzen weiter fortpflanzen. Schätzungen gehen daher davon aus, daß 10 bis 30% aller Genom-Annotationen falsch sein könnten [62, 63].

Falsch annotierte Enzyme können die spätere Auswertung der erhaltenen Einteilung erheblich erschweren, da durch sie eine mechanistische Verwandtschaft zwischen Enzymfunktionen impliziert wird, die in Wirklichkeit nicht besteht. Daher werden, soweit möglich, nur solche Enzymsequenzen in den Datensatz integriert, deren funktionelle Klassifizierung auf experimentellem Wege ermittelt wurde. Da die Datenbanken über keine entsprechenden Einträge verfügen, erfolgt die Auswahl mittels Umkehrschluß. Alle Enzymeinträge, deren Beschreibung aufgrund bestimmter Schlüsselbegriffe (vgl. Abschnitt 2.3.3) darauf schließen läßt, daß es sich lediglich um eine hypothetische Zuordnung handelt, werden vom Datensatz ausgeschlossen.

Die Datenbanken enthalten überdies eine Vielzahl von fragmentarisch sequenzierten Enzymen. Da diese oftmals nur einen Teil der katalytischen Domäne beinhalten, wichtige Sequenzabschnitte indes fehlen, gilt ergänzend ein Längenkriterium, das fragmentarische Enzymsequenzen ausschließt, bzw. eine Mindestlänge der Sequenz von 100 Aminosäuren garantiert.

Letztlich verbleiben so 62.875 Enzymsequenzen, die alle geforderten Kriterien erfüllen. Dies entspricht etwa 43,6% der insgesamt identifizierten Enzymsequenzen bzw. einem Anteil von ca. 5,2% am Gesamtdatenbestand beider Datenbanken.

3.2.2 Aktualisierung der EC-Klassifikation

Die EC-Klassifikation ist keine starre Einteilung, sondern unterliegt ständigen Änderungen und Aktualisierungen. Fortlaufend kommen neue EC-Klassen hinzu, während bestehende Einträge gelöscht, transferiert oder zusammengefaßt werden. Infolgedessen enthalten Swiss-Prot und TrEMBL zahlreiche Sequenzeinträge, deren Zuordnung nicht mehr der aktuellen EC-Klassifikation entsprechen. Dies kann dazu führen, daß während der Clusterung vermeintlich Enzyme unterschiedlicher Funktion miteinander assoziiert werden, obwohl diese mittlerweile derselben Enzymklasse angehören. Umgekehrt ist es möglich, daß Sequenzen nur scheinbar zur selben EC-Klasse gehören, mittlerweile jedoch unterschiedlichen Funktionen zugeordnet sind. Um den daraus resultie-

renden Problemen bei der Auswertung des Clusterergebnisses vorzubeugen und sicherzustellen, daß die verwendete Einteilung tatsächlich der derzeit aktuellen entspricht, werden alle im Datensatz vorkommenden EC-Klassen auf ihre Gültigkeit geprüft und gegebenenfalls aktualisiert.

Im hier beschriebenen Datensatz wurden 85 EC-Klassen mit insgesamt 1.158 Sequenzeinträgen aktualisiert. Zwei der EC-Klassen wurden von der IUBMB ersatzlos gestrichen. Die ihnen zugeordneten Sequenzen wurden dementsprechend aus dem Datensatz entfernt (vgl. Tabelle 3-22). Die Sequenzen dreier Enzymklassen wurden auf mehrere unterschiedliche EC-Klassen verteilt. Da nicht festgestellt werden kann, welcher neuen Enzymklasse eine jeweilige Sequenz nun angehört, wurde der ursprünglichen EC-Nummer lediglich eine führende Null zur Kennzeichnung vorangestellt. Ihre endgültige Zuordnung ergibt sich dann nach Abschluß der Clusterung. Alle übrigen EC-Klassen wurden entsprechend der Vorgaben der IUBMB aktualisiert.

Tabelle 3-1: Während der Aktualisierung der EC-Klassifikation gelöschte bzw. nicht einheitlich transferierte Enzymklassen

EC-Klasse	Status	betroffene Sequenzen
1.12.99.2	gelöscht	1
3.2.1.19	gelöscht	1
2.1.1.24	jetzt EC 2.1.1.77, 2.1.1.80 oder 2.1.1.100	2
3.4.21.14	jetzt EC 3.4.21.62, EC 3.4.21.65 oder EC 3.4.21.67	3
3.4.22.17	jetzt EC 3.4.22.52 oder EC 3.4.22.53	4

3.2.3 Struktur des Enzymdatensatzes

Nach Berücksichtigung aller Änderungen im EC-System umfaßt die endgültige Fassung des hier beschriebenen Datensatzes 62.873 Sequenzen klassifizierter Enzyme. Diese gehören 2.143 verschiedenen EC-Klassen an. Für die übrigen der insgesamt etwa 3.800 derzeit gültigen Enzymklassen sind keine Sequenzen im Datensatz enthalten.

Abbildung 3-1a zeigt die prozentuale Verteilung der Sequenzen auf die sechs Hauptklassen der Enzyme. Etwa ¾ aller Sequenzen entfallen auf die ersten drei Hauptklassen der Oxidoreduktasen, Transferasen und Hydrolasen. Lediglich

24% der Enzymsequenzen gehören den Lyasen, Isomerasen und Ligasen an. Die genauen Zahlenverhältnisse sind in Tabelle 3-2 aufgeführt.

In Abbildung 3-1b ist für jede Hauptklasse die Anzahl der im Datensatz enthaltenen EC-Klassen dargestellt. Die sequenzreichsten Hauptklassen sind zugleich auch mit den meisten Enzymklassen vertreten. Tabelle 3-2 gibt eine Aufgliederung der Zahlenwerte entsprechend der Ebenen der EC-Klassifikation wieder.

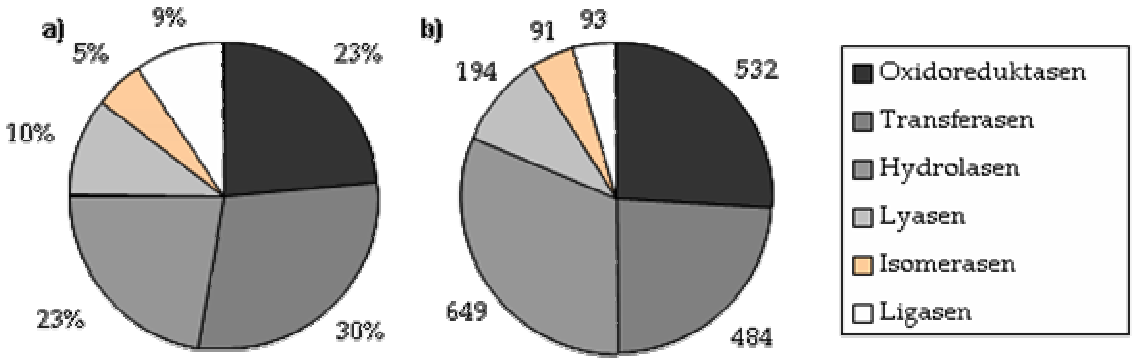


Abbildung 3-1: a) Prozentuale Verteilung der Sequenzen auf die sechs Hauptklassen des EC-Systems b) Anzahl der im Datensatz erhaltenen EC-Klassen für jede der sechs Hauptklassen

Im Durchschnitt entfallen etwa 29 Sequenzen auf eine Enzymklasse. Eine genaue Betrachtung der Sequenzzahlen zeigt jedoch, daß für den überwiegenden Teil der EC-Klassen nur sehr wenige Sequenzen verfügbar sind. 343 (16%) der insgesamt 2.143 vertretenen Enzymklassen sind lediglich mit einer einzelnen Sequenz im Datensatz vertreten. Etwa 44% der EC-Klassen umfassen zwischen 2 und 10 Sequenzen. In vereinzelten Fällen enthält der Datensatz aber auch mehrere hundert Sequenzen zu einer einzelnen Enzymklasse. So ist beispielsweise die sequenzreichste Enzymfamilie der NADH-Ubichinon Oxidoreduktasen (EC 1.6.5.3) mit insgesamt 2140 Sequenzen im Datensatz vertreten.

1511 Enzyme des Datensatzes sind multifunktional und werden daher gleich mehreren Enzymklassen zugerechnet. So ist beispielsweise die menschliche Fettsäure-Synthase (Swiss-Prot-ID: FAS_HUMAN; P49327) unter insgesamt acht verschiedenen EC-Klassen verzeichnet.

Tabelle 3-2: Verteilung der Enzymsequenzen auf die sechs Hauptklassen des EC-Systems und Anzahl der unterschiedlichen EC-Klassen für jede Ebene der Klassifikation. Die abweichende Gesamtzahl der Sequenzen ist auf multifunktionale Enzyme zurückzuführen, die mehreren EC-Klassen angehören.

Enzymklasse		Sequenzen	Subklassen	Sub-Subklassen	EC-Klassen
1	Oxidoreduktasen	15.167	21	85	532
2	Transferasen	18.729	9	26	585
3	Hydrolasen	14.718	11	48	649
4	Lyasen	6.744	7	15	193
5	Isomerasen	3.318	6	16	91
6	Ligasen	6.031	6	10	93
	Insgesamt	**64.707**	**60**	**200**	**2.143**

3.3 Das Sequenzalignment

3.3.1 Sequenz- vs. Strukturvergleich

Homologe Proteine gehen auf einen gemeinsamen „Vorfahr“ zurück, aus dem sie durch divergente Evolution hervorgegangen sind [8] und besitzen infolgedessen eine ähnliche räumliche Struktur und oftmals auch eine nachweisbare Sequenzähnlichkeit [7]. Dies kann in Struktur- und Sequenzvergleichen genutzt werden, um homologe Proteine zu identifizieren. Strukturvergleiche sind insbesondere geeignet, um auch entfernte Homologien aufzudecken [26], da die Struktur eines Proteins evolutionär stärker konserviert ist als seine Sequenz [16]. Die Ursache hierfür besteht darin, daß die Funktion eines Proteins im wesentlichen eine bestimmte Struktur erfordert, diese aber über verschiedene Sequenzen realisiert werden kann. So findet sich strukturelle Ähnlichkeit auch in Fällen geringer Sequenzidentität [64, 65]. Die experimentelle Aufklärung von Proteinstrukturen mittels Röntgenstrukturanalyse und NMR ist jedoch nur mit großem Aufwand möglich, während die Sequenzierung von Protein- und DNA-Sequenzen vergleichsweise einfach durchzuführen ist. Infolgedessen ist die Zahl bekannter Sequenzen weitaus größer als die Anzahl aufgeklärter Proteinstrukturen. In der vorliegenden Arbeit wird daher die Sequenzähnlichkeit zur Detektierung von Homologen verwendet. Die Empfindlichkeit dieser Methode ist jedoch aus oben genannten Gründen begrenzt. Geringe Sequenzübereinstimmungen lassen überdies viel Raum für Interpretationen.

3.3.2 Statistische Signifikanz der Sequenzalignments

Da der gemeinsame „Vorfahr“ homologer Proteine in der Regel nicht bekannt ist, wird von Homologie ausgegangen, wenn eine statistisch signifikante Sequenzähnlichkeit vorliegt. Ähnlichkeit ist im Gegensatz zu Homologie jedoch nicht transitiv [7]. Sequenzähnlichkeit kann quantifiziert werden, während Homologie eine Beziehung ist, die entweder besteht oder nicht besteht [9]. Ähnliche Sequenzen implizieren daher nicht notwendigerweise eine evolutionäre Verwandtschaft der Proteine. Aus diesem Grund erfordert ein Alignment normalerweise stets eine kritische Betrachtung durch den Anwender, um sicherzustellen, daß die ermittelte Sequenzübereinstimmung tatsächlich signifikant ist und infolgedessen auf eine Homologie der Sequenzen schließen läßt. Die große Anzahl der zu untersuchenden Sequenzen macht eine manuelle Beurteilung der Alignments jedoch unmöglich. Daher kommt der Wahl eines geeigneten Maßes sowie eines geeigneten Grenzwertes zur automatischen Bewertung der Alignments eine besondere Bedeutung zu.

3.3.2.1 Der *E-value* als Maß der Sequenzähnlichkeit

BLAST [13, 40] errechnet zu jedem Sequenzvergleich vier Werte, anhand derer die Qualität eines Alignments eingeschätzt werden kann (vgl. Abschnitt 2.4.2.2). Die prozentuale Identität der Sequenzen ist – obwohl häufig zur Quantifizierung der Sequenzübereinstimmung verwendet – kaum geeignet, um Sequenzen bezüglich ihrer Ähnlichkeit zu bewerten [66], da sie keine Information über die Länge der miteinander alignierten Sequenzbereiche enthält. Infolgedessen erscheint beispielsweise eine 50%ige Sequenzidentität über nur wenige Aminosäuren signifikanter als eine 40%ige Identität, die sich über die gesamte Länge der Sequenz erstreckt. Am effektivsten erfolgt eine Beurteilung anhand des statistischen *scores* oder des *E-values* [66]. In der vorliegenden Arbeit wurde der *E-value* verwendet, um den Grad der Identität zweier Sequenzen zu quantifizieren, da dieser sowohl den für das Alignment ermittelten *score* als auch dessen statistische Signifikanz berücksichtigt. Die Verwendung des *E-values* erfordert jedoch eine entsprechend angepaßte Vorgehensweise.

Der *E-value* ergibt sich, wie in Abschnitt 2.4.2.2 beschrieben, aus der Gleichung:

$$E = \frac{N}{2^{S_n}} \qquad (3.1)$$

S_n entspricht dem normalisierten *score* (*bit score*) des Alignments. N ergibt sich als Produkt aus m und n, wobei m die Länge der Suchsequenz und n die Größe des Sequenzraumes ist [40]. Wird eine Sequenz gegen eine Sequenzdatenbank verglichen, entspricht n der Größe der Datenbank (Summe der Aminosäuren aller enthaltenen Sequenzen). Da nicht die Länge der jeweils mit der Suchsequenz verglichenen Datenbanksequenz, sondern die Gesamtlänge der Datenbank zur Berechnung des *E-values* verwendet wird, ist der Wert des *E-values* indirekt von der Richtung abhängig, in der das Alignment ausgeführt wird. Werden zwei sehr unterschiedlich lange Sequenzen A und B miteinander verglichen, so kann der *E-value*, je nach Ausrichtung des Vergleichs (A → B oder B → A), um mehrere Zehnerpotenzen differieren. Da anhand des *E-values* jedoch auf eine mögliche Homologie der Sequenzen geschlossen werden soll, würde dies im Grenzbereich dazu führen, daß die Sequenzen, je nach Richtung des Alignments mal als homolog und mal als nicht-homolog eingestuft würden. Um sicherzustellen, daß die verwendeten Alignments den geforderten Grenzwert zweifelsfrei erfüllen (vgl. Abschnitt 3.3.2.2), werden daher alle Sequenzvergleiche ausschließlich in der Orientierung ausgeführt, die den höheren und somit weniger signifikanten *E-value* erzielt. Liegt dieser dennoch innerhalb der gesetzten Grenzen, so kann dies mit einiger Sicherheit als Beleg für eine gemeinsame Abstammung angesehen werden.

Die gerichteten Sequenzvergleiche werden erzeugt, indem die jeweilige Suchsequenz (*Query*) ausschließlich gegen Datenbanksequenzen gleicher oder geringer Sequenzlänge verglichen wird. Dies hat jedoch eine ständige Größenänderung der zum Vergleich verwendeten Datenbank zur Folge, welches sich wiederum auf die errechneten *E-values* auswirkt (s. o.). Um trotz wechselnder Datenbankgröße miteinander vergleichbare *E-values* zu erhalten, muß daher vor Ausführung der Alignments die tatsächliche Gesamtgröße der Datenbank ermittelt und als Parameter an BLAST übergeben werden, so daß für alle *E-values* die gleiche Berechnungsgrundlage gegeben ist.

Die Beschränkung auf nur eine Alignmentrichtung bringt überdies eine erhebliche Reduktion der Zahl auszuführender Sequenzvergleiche mit sich. Wird, wie bei einem „alle vs. alle"-Alignment üblich, der Datensatz vollständig gegen sich selbst verglichen, so ergibt sich die Zahl der auszuführenden Alignments aus x^2, wobei x der Anzahl der Sequenzen entspricht. Beschränkt man sich indes auf nur eine Richtung, verbleiben lediglich $(x*(x+1))/2$ Vergleiche[4]. Für den hier beschriebenen, aus 62.873 Sequenzen bestehend Sequenzdatensatz ergibt sich somit für beide „alle vs. alle"-Alignments eine Reduktion von ursprünglich 7,9 Milliarden auf etwa 3,95 Milliarden Sequenzvergleiche. Die benötigte Rechenzeit kann dadurch von ursprünglich vier auf 2 ½ Tage reduziert werden (Angabe bezieht sich auf eine SunFire V880 unter gleichzeitiger Verwendung vier der insgesamt sechs UltraSparc III/750 Prozessoren).

3.3.2.2 Bestimmung eines Grenzwertes

Es läßt sich nur schwer ein klarer Grenzwert festlegen, um zuverlässig zwischen echter und zufälliger Sequenzähnlichkeit zu unterscheiden. Die Wahl eines zu rigorosen Grenzwertes führt dazu, daß wichtige Ähnlichkeiten unberücksichtigt bleiben, während ein zu permissiver Grenzwert falsche Verbindungen erzeugt. *E-values* kleiner 10^{-3} können mit einiger Sicherheit als signifikant angesehen werden, während Werte über 10 im allgemeinen zufällige Ähnlichkeiten reflektieren [66]. Die Signifikanz der *scores* im mittleren Wertebereich läßt sich indes nur schwer einschätzen.

Yona *et al*. (1998) konnten in ihrer Studie zeigen, daß für BLASTP ein *E-value* von 10^{-3} noch auf eine statistisch signifikante Ähnlichkeit der Proteine schließen läßt. Dieser Wert wurde anhand der Verteilung aller *E-values* bestimmt, die sich für die Suche einer Sequenz über die komplette Swiss-Prot ergab. Da der hier verwendete Datensatz jedoch nur eine Auswahl aller in Swiss-Prot und TrEMBL enthaltenen Sequenzen repräsentiert, und BLAST seitdem mehrere Versionsänderungen erfahren hat, wurde die Gültigkeit des Wertes nochmals

[4] Alignments, bei denen die Sequenz mit sich selbst verglichen wird, werden beibehalten. Der so erhaltene Wert für den *E-value* stellt aufgrund der 100%igen Identität der miteinander verglichenen Sequenzen den niedrigsten Wert dar, der im Vergleich mit der jeweiligen Suchsequenz erzielt werden kann und ermöglicht somit eine zusätzliche Einschätzung der Signifikanz der übrigen Alignmenttreffer.

anhand des eigenen Datensatzes überprüft. Abbildung 3-2 zeigt die für den Sequenzdatensatz ermittelte Verteilung der *E-values*.

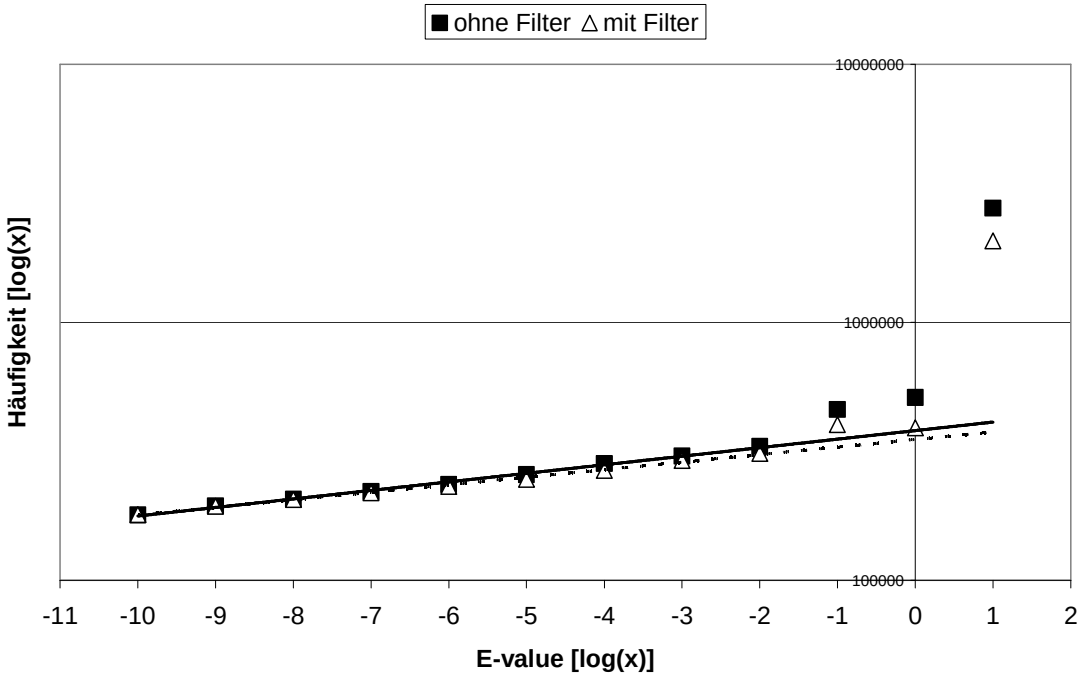


Abbildung 3-2: Gesamtverteilung aller für ein „alle vs. alle"-Alignment ermittelten *E-values* für den Bereich von 10^{-10} bis 10. Die Verteilung der Werte ist in doppelt-logarithmischer Auftragung dargestellt. Die quadratischen Datenpunkte (■) repräsentieren den Verlauf für das ungefilterte „alle vs. alle"-Alignment, während die dreieckigen Datenpunkte (Δ) das Ergebnis des gefilterten Alignmentlaufes kennzeichnen.

Die Änderungsrate von Aminosäuresequenzen variiert stark von Protein zu Protein und beeinflußt folglich, wieweit der Ursprung eines jeweiligen Proteins zurückverfolgt werden kann. Demzufolge fällt auch die bei einer Datenbanksuche ermittelte Verteilung der *E-values* für jede Sequenz unterschiedlich aus. Um einen eindeutigen Grenzwert zu erhalten ist es daher notwendig, die Einzelverteilungen zu mitteln. Aus diesem Grund wurde die Gesamtverteilung aller *E-values* für ein jeweiliges „alle vs. alle"-Alignment verwendet, um einen geeigneten Grenzwert zu bestimmen.

Bei doppelt-logarithmischer Auftragung ist die Verteilung der niedrigen *E-values* sowohl für das ungefilterte (durchgezogene Linie) als auch das gefilterte „alle vs. alle"-Alignment (unterbrochene Linie) annähernd linear, steigt ab einem Wert von 10^{-1} jedoch rapide an (vgl. Abbildung 3-2). Yona *et al*. zufolge

beruht die zunehmende Steigung bei höheren *E-values* auf einer steigenden Zahl von Alignments mit Sequenzen, die nicht mit der Suchsequenz verwandt sind. Daher wird der Wert als Grenzwert gewählt, bei dem sich die Steigung zu ändern beginnt. Dementsprechend werden im folgenden nur solche Alignments berücksichtigt, deren *E-values* kleiner 10^{-1} sind.

3.3.3 Vermeidung systematischer Fehler bei der Detektierung von Homologen

3.3.3.1 Regionen sehr kurzer Sequenzübereinstimmung

Eine erste Begutachtung der erhaltenen Sequenzalignments ergab, daß trotz des gewählten Grenzwertes Alignments von sehr geringer Länge entstehen können. Die kürzesten unter ihnen umfassen lediglich 16 Aminosäuren. Aufgrund des geringen Informationsgehalts können derartige Sequenzübereinstimmung nicht als signifikant angesehen werden. Die Ähnlichkeit homologer Sequenzen erstreckt sich in der Regel über die gesamte Länge einer Sequenz oder Domäne. Eine zufällige Identität von mehr als 50% über eine Region von 20 bis 40 Aminosäuren ist hingegen recht häufig [6] und stellt folglich kein verläßliches Indiz für Homologie dar.

Da der *E-value* als alleiniges Kriterium nicht ausreicht, um derartige Alignments auszuschließen, wird zusätzlich ein Längenkriterium verwendet. Dessen Länge orientiert sich an der kürzesten im Datensatz definierten Domäne. Die sogenannte SH3-Domäne ist Bestandteil vieler intrazellularer und membrangebundener Proteine und findet sich sowohl in Enzymsequenzen als auch in Proteinen des Zytoskeletts. Ihre durchschnittliche Länge beträgt etwa 60 Aminosäuren [67]. Dieser Wert wurde somit als Mindestlänge eines Alignments definiert. Zufällige Ähnlichkeiten dieser Länge erscheinen eher unwahrscheinlich und können daher mit einiger Sicherheit als signifikant angesehen werden.

3.3.3.2 Regionen ungewöhnlicher Aminosäurezusammensetzung

Eine ungewöhnliche Aminosäurezusammensetzung der Suchsequenz kann das Resultat einer Suche stark beeinflussen. Die Statistik der Datenbanksuche basiert auf der Annahme, daß nicht-homologe Proteine keine Sequenzähnlichkeit

aufweisen. Bestimmte Sequenzmuster widersprechen jedoch dieser Annahme. Ein typischer Fall sind Sequenzbereiche geringer kompositorischer Komplexität (*low complexity regions*, kurz LCRs) [68]. LCRs sind Sequenzbereiche, die zum Beispiel aus langen Wiederholungen eines einzelnen oder einiger weniger Aminosäuren bestehen (z. B. Poly-Alanin-Folgen, Prolin-reiche Regionen) [44]. Solche lokalen Bereiche sehr einfacher Aminosäurezusammensetzung sind sehr verbreitet. Mehr als die Hälfte aller Sequenzen in den Datenbanken enthalten mindestens eine solche Region [69].

Die standardmäßig verwendeten Alignment- und Bewertungsverfahren sind nicht darauf ausgelegt, die evolutionären Prozesse, die zur Entstehung dieser Regionen geführt haben, entsprechend zu berücksichtigen. Infolgedessen können Sequenzen, die in diesen Regionen übereinstimmen, sehr hohe Ähnlichkeitswerte erreichen. Es ist jedoch unwahrscheinlich, daß Alignments, deren Ähnlichkeiten allein auf wiederholte Sequenzbereiche basieren, eine signifikante Übereinstimmung zwischen den Sequenzen reflektieren. Um solche statistisch signifikanten, aber evolutionär wenig informativen Sequenzübereinstimmungen zu vermeiden, wird für eines der „alle vs. alle"-Alignments der in BLASTP integrierte Sequenzfilter SEG [43, 44] verwendet. Dieser maskiert die entsprechenden Abschnitte der Suchsequenz, so daß diese bei einem Sequenzvergleich unberücksichtigt bleiben. Auf diese Weise können Sequenzpaarungen, deren Ähnlichkeit ausschließlich auf Sequenzbereiche geringer kompositorischer Komplexität beruht, vermieden werden.

Aufgrund ihrer Verbreitung können LCRs aber auch integrale Bestandteile tatsächlich homologer Sequenzabschnitte sein. Da diese in der Regel aber eine über die LCRs hinausgehende Ähnlichkeit besitzen, verhindert die Verwendung des Sequenzfilters in diesen Fällen das Zustandekommen eines Alignments nicht. Es beeinflußt jedoch mitunter dessen Resultat.

Liegt ein LCR innerhalb des homologen Sequenzbereichs, so kann dessen Maskierung die Länge des resultierenden Alignments und infolgedessen auch den für das Alignment errechneten *E-value* beeinflussen. Ist der mittels Filter maskierte Bereich beispielsweise N- oder C-terminal im homologen Sequenzabschnitt lokalisiert, so ist ein entsprechend verkürztes Alignment die Folge. Liegt er hingegen inmitten des homologen Sequenzabschnitts, hängt das Ergebnis von seiner Länge ab. Ist die maskierte Region nur kurz, so kann sie von BLASTP durch Einfügen von *gaps* überbrückt werden. Übersteigt sie jedoch ei-

ne gewisse Länge, so kann dies zu einem vorzeitigen Abbruch während der Verlängerung des Alignments führen (vgl. Abschnitt 2.4.2.1), in dessen Folge dann zwei separate Alignments erzeugt werden.

Da mittels der Alignments aber nicht nur auf eine mögliche Homologie der Sequenzen sondern anhand der erhaltenen Alignmentgrenzen auch auf die modulare Domänenarchitektur der Proteine geschlossen werden soll, ist dieser Effekt der Filterung unerwünscht. Aus diesem Grund wird das „alle vs. alle"-Alignment nochmals ohne Einsatz des Sequenzfilters wiederholt. Die so erhaltenen lokalen Alignments und ihre *E-values* dienen schließlich als Grundlage für alle weiteren Schritte. Es werden jedoch nur die Sequenzpaarungen des ungefilterten Alignmentlaufes verwendet, die auch bei Verwendung des Sequenzfilters zustande kommen. Tabelle 3-3 gibt die Anzahl der resultierenden Sequenzpaarungen für beide Alignmentläufe wieder. Die Differenz beider Werte entspricht der Anzahl von Sequenzpaarungen, deren Ähnlichkeiten ausschließlich auf Bereiche geringer Komplexität beruhen und deshalb keine Verwendung fanden.

Tabelle 3-3: Anzahl der erhaltenen Sequenzpaarungen für das mit SEG gefilterte und das ungefilterte „alle vs. alle"-Alignment

Alignmentlauf	**Zahl der Sequenzpaarungen**
ohne Sequenzfilter	6.086.971
mit Sequenzfilter (SEG)	5.915.568
Differenz	**171.403**

3.4 Die Domänenstrukturvorhersage

3.4.1 Problematik bei der Klassifizierung modularer Enzyme

Die Suche nach homologen Sequenzen mittels BLAST oder ähnlicher Programme kann zu unterschiedlichen Resultaten führen. Im günstigsten Fall erbringt die Datenbanksuche eine klare Ähnlichkeit zu Proteinen einer Funktion, die sich über die gesamte Länge der Sequenz erstreckt. Das weitaus häufigere Resultat ist jedoch eine Liste von Treffern mit partieller Sequenzähnlichkeit zu verschiedenen Proteinen mit zum Teil unterschiedlicher Funktion.

Partielle Sequenzähnlichkeit ist meist die Folge einer modularen Proteinarchitektur. Die Proteine eukaryotischer Organismen bestehen oftmals aus Kombinationen mehrerer autonomer Module bzw. Domänen [31, 32]. „Domäne“ ist in diesem Zusammenhang nicht gleichbedeutend mit einer Strukturdomäne, sondern bezeichnet treffender eine Sequenzdomäne. Der Begriff „Strukturdomäne“ umschreibt einen sich unabhängig von der übrigen Proteinsequenz zu einer Tertiärstruktur faltenden Sequenzabschnitt. Der Begriff „Sequenzdomäne“ bezeichnet hingegen den Abschnitt einer Aminosäuresequenz, der signifikante Ähnlichkeit zu Abschnitten anderer Proteine besitzt. Sequenzdomänen werden daher auch als Homologiedomänen bezeichnet. Eine Sequenzdomäne kann einer Strukturdomäne entsprechen, aber es gibt bislang keine Studie darüber, in wie weit Sequenzdomänen tatsächlich mit Strukturdomänen korrelieren.

Manche Sequenzdomänen können als evolutionär mobil angesehen werden, da sie in Proteinen gleicher Domänenzusammensetzung in unterschiedlicher Reihenfolge, oder aber in Proteinen unterschiedlicher Domänenzusammensetzung in Kombination mit ganz verschiedenen Domänen auftreten können [31, 33]. Infolgedessen besitzen Sequenzen, die aus mehreren evolutionär mobilen Domänen zusammengesetzt sind, in unterschiedlichen Sequenzbereichen Ähnlichkeit zu verschiedenen Proteinfamilien. Solche Proteine sind daher mehreren Clustern gleichzeitig zuzuordnen. Hierarchische Cluster-Verfahren erzeugen jedoch Hierarchien disjunkter Cluster. Ein Verfahren, das disjunkt klassifiziert, liefert "scharf getrennte" Cluster, bei dem jedes Objekt genau einem Cluster angehört. Eine umfassende Klassifizierung der Multidomänenenzyme ist auf dieser Grundlage nicht möglich, und birgt überdies die Gefahr, Enzymfunktionen irrtümlich miteinander zu assoziieren:

- Besitzen zwei Enzyme nur partielle Sequenzähnlichkeit, so kann sich diese auf jede der enthaltenen Domänen beziehen. Dies schließt auch jene ein, die nicht unmittelbar mit der eigentlichen Funktion assoziiert sind, wie zum Beispiel gemeinsame Signal-, Bindungs- oder regulatorische Domänen. Eine Gruppierung der Sequenzen allein aufgrund von Ähnlichkeit zu einer dieser Domänen würde eine mechanistische Verwandtschaft zwischen den Enzymfunktionen implizieren, die in Wirklichkeit nicht besteht.

- Manche eukaryotischen Multidomänenenzyme sind überdies Multienzymkomplexe, in denen verschiedene Enzyme kovalent zu einer einzigen Polypeptidkette verknüpft sind. Diese multifunktionalen Enzyme arbeiten mittels verschiedener Mechanismen. Eine Gruppierung der vollständigen Sequenz würde ein gemeinsames Auftreten von Enzymaktivitäten innerhalb eines Clusters implizieren, das tatsächlich nicht vorhanden ist.
- Ein weiteres Problem ergibt sich bei Verfahren, die über Intermediärsequenzen nach homologen Proteinen suchen, wie dies beim verwendeten transitiven *single linkage* Verfahren der Fall ist. So ist es möglich zwei nicht homologe Sequenzen über eine Intermediärsequenz miteinander zu gruppieren, die je eine Domäne mit jeder der Sequenzen teilt (vgl. Abbildung 3-3). Doch nur wenn sich die Sequenzen innerhalb eines Clusters in der gleichen Region ähneln, kann dies zur Einsichten bezüglich der Verwandtschaft zwischen den verschiedenen Enzymklassen führen.

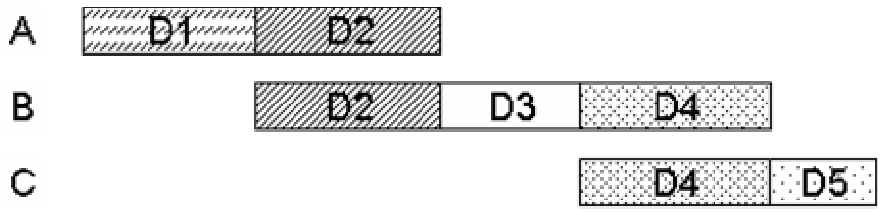


Abbildung 3-3: A, B und C sind drei Multidomänensequenzen, die aus den Domänen D1 bis D5 bestehen. Da die Domänen D2 und D4 in Sequenz B enthalten sind, werden die Sequenzen A und C miteinander gruppiert, obwohl sie keine gemeinsame Domäne besitzen.

Aus diesen Gründen ist die Berücksichtigung der modularen Proteinarchitektur eine grundlegende Voraussetzung, um die Enzyme korrekt zu klassifizieren. Nur so können alle Verwandtschaftsbeziehungen einbezogen und zugleich irreführende Verknüpfungen vermieden werden. Für einen Teil der Enzymsequenzen ist bisher jedoch noch keine Domänenstruktur beschrieben worden. Die Aufgabe bestand daher zunächst darin, entsprechende Daten für alle im Datensatz enthaltenen Sequenzen zu generieren.

3.4.2 Methoden zur Ermittlung der modularen Proteinarchitektur

Die Identifizierung der Domänenstruktur einer Sequenz kann auf verschiedene Weise erfolgen. Methoden, wie sie von den Datenbanken Dali Domain Dictionary (DDD) [70], CATH [71] und SCOP [72] verwendet werden, nutzen strukturelle Daten um Domänen zu lokalisieren und zuzuordnen. Doch selbst bei Verwendung struktureller Daten ist die vollständige Automatisierung der Domänenzuordnung kein triviales Problem [73] und erfordert überdies eine aufgeklärte Struktur.

Die Identifizierung der Domänen auf Sequenzebene beruht häufig auf der Entdeckung globaler/lokaler Sequenzähnlichkeit zwischen der untersuchten Sequenz und Domänensequenzen in Datenbanken wie zum Beispiel Pfam [46]. Es ergeben sich jedoch Schwierigkeiten bei der Aufklärung der Domänenstruktur, wenn die Suche gegen die Datenbank keine signifikante Ähnlichkeit zu bekannten Domänen erbringt. In solchen Fällen wird ein *ab initio* Verfahren für die Domänenstrukturvorhersage benötigt.

Bisher existieren nur wenige Verfahren, um Domänen allein anhand von Sequenzinformationen vorhersagen. Diese sind zudem nur selten vollständig automatisiert. Sie basieren im wesentlichen auf der Suche nach Homologen (z. B. mittels iterativer PSI-BLAST Suchen (ProDom) [74] oder *Hidden Markov Model* Profilen (Pfam) [46]) und anschließenden multiplen Sequenzalignments. Während das optimale Alignment von zwei Sequenzen mittels eines Computers recht schnell berechnet werden kann, ist dies beim multiplen Sequenzalignment nicht mehr ohne weiteres möglich, da die Komplexität mit der Anzahl der Sequenzen drastisch zunimmt. Daher werden in der Regel Heuristiken verwendet, die beispielsweise zuerst einen phylogenetischen Baum berechnen, und davon ausgehend jeweils paarweise Alignments durchführen. Nichtsdestotrotz bleiben multiple Sequenzalignments zeit- und rechenintensiv und sind daher auf relativ geringe Sequenzzahlen begrenzt. Im vorliegenden Fall gilt es jedoch, während eines Programmlaufes die Domänenstruktur von mehreren zehntausend Sequenzen zu ermitteln. Um eine ausreichende Geschwindigkeit zu erzielen wurde daher eine eigene Methode entwickelt, um für eine gegebene Proteinsequenz putative Sequenzdomänen und ihre Grenzen vorherzusagen.

3.4.3 Domänenstrukturvorhersage mittels lokaler Sequenzalignments

Bei der verwendeten Methode handelt sich um eine intuitive Prozedur, die von der Prämisse ausgeht, daß sich die Ähnlichkeit homologer Sequenzen üblicherweise über die gesamte Länge einer Sequenz oder Domäne erstreckt. Demzufolge sollte sich anhand der relativen Positionen der lokalen Alignments zur Suchsequenz die Lage und Ausdehnung der einzelnen Domänen bestimmen lassen. Dies soll im folgenden anhand eines Multidomänenenzyms bekannter Domänenstruktur illustriert werden:

Das multifunktionale AROM-Enzym aus *Saccharomyces cerevisiae* (Swiss-Prot-ID: ARO1_YEAST; P08566) besteht aus fünf monofunktionalen Domänen [75]. Das Enzymsystem arbeitet sequentiell indem es fünf aufeinanderfolgende, durch gemeinsame metabolische Intermediate miteinander verbundene Reaktionen in der prechorismaten polyaromatischen Aminosäurebiosynthese katalysiert (vgl. Tabelle 3-4). Der Vorteil einer solchen Anordnung besteht darin, daß die Synthese der einzelnen Enzyme koordiniert ist. Zudem können instabile Zwischenprodukte geschützt und effizient von einem aktiven Zentrum zum nächsten weitergeleitet werden, ohne das Aggregat zu verlassen. Damit werden die Diffusionszeiten erheblich verkürzt und Nebenreaktionen minimiert.

Tabelle 3-4: Das multifunktionale AROM-Enzym aus *Saccharomyces cerevisiae* katalysiert fünf aufeinanderfolgende, durch gemeinsame metabolische Intermediate miteinander verbundene Reaktionen.

Domäne	Reaktion
3-Dehydroquinat Synthase (EC 4.2.3.4)	3-Deoxy-arabino-heptulonat-7-phosphat = 3-Dehydroquinat + Phosphat
3-Dehydroquinat Dehydratase (EC 4.2.1.10)	3-Dehydroquinat = 3-Dehydroshikimat + H_2O
Shikimat Dehydrogenase (EC 1.1.1.25)	3-Dehydroshikimat + NADPH + H^+ = Shikimat + $NADP^+$
Shikimat Kinase (EC 2.7.1.71)	Shikimat + ATP = Shikimat-3-phosphat + ADP
3-Phosphoshikimat 1-carboxyvinyltransferase (EC 2.5.1.19)	Shikimat-3-phosphat + Phosphoenol-pyruvate = Phosphat + 5-O-(1-Carboxyvinyl)-3-phosphoshikimat

Das Verfahren zur Vorhersage der Domänengrenzen verwendet einen Algorithmus, um diejenigen Aminosäurepositionen innerhalb der Suchsequenz zu identifizieren, die mit den N- oder C-Termini der Datenbanktreffer aligniert wurden. Zudem wird für jede Position die Häufigkeit ermittelt, mit der sie als N- bzw. C-Terminus dient. Sequenzpositionen, die sowohl als N- als auch als C-Terminus lokaler Alignments fungieren, haben eine erhöhte Wahrscheinlichkeit, die Grenzen zwischen zwei Domänen zu markieren und erhalten daher eine höhere Bewertung, die sich als Summe der N- und C- terminalen Häufigkeitswerte ergibt.

Für das hier betrachtete AROM-Enzym ergeben sich während der „alle vs. alle“-Sequenzvergleiche 429 Alignments, die alle geforderten Kriterien erfüllen (vgl. Abschnitt 2.4.2.3). Abbildung 3-4 zeigt eine graphische Repräsentation ihrer N- und C-Termini entlang der Sequenz (vgl. Abschnitt 2.5).

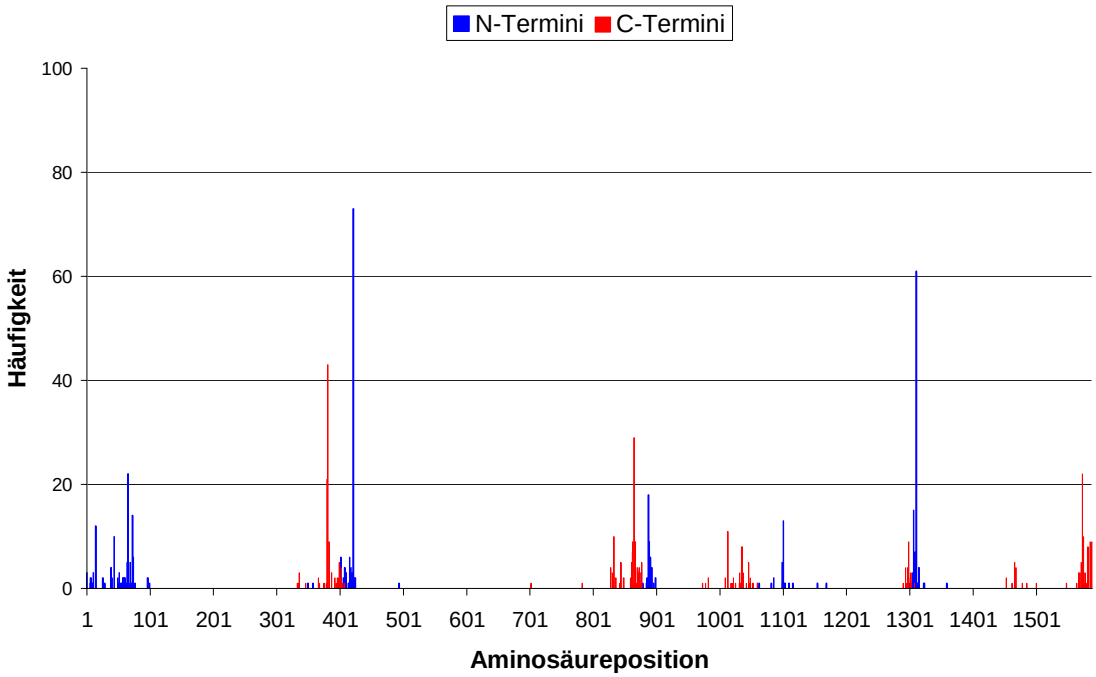


Abbildung 3-4: Graphische Repräsentation der N- (blau) und C-Termini (rot) aller 429 zur Swiss-Prot Sequenz ARO1_YEAST verfügbaren lokalen Sequenzalignments.

Die graphische Darstellung veranschaulicht, daß sich die N- und C-Termini der Alignments nicht zufällig über die Sequenz verteilen, sondern sich der Hypothese entsprechend auf bestimmte Sequenzbereiche konzentrieren. Dazwischen liegen wiederum Sequenzabschnitte, in denen sich keine oder nur sehr wenige Alignment-Termini befinden. Eine farbliche Unterscheidung der verschiedenen Termini offenbart, daß diese Abschnitte jeweils von N- und C-Termini flankiert werden. Daraus läßt sich ableiten, daß sich das Enzym in insgesamt fünf Sequenzdomänen gliedert.

Der nächste Schritt besteht darin, die Grenzen der Sequenzdomänen zu ermitteln. Obwohl sich die grundlegende Struktur des Enzyms bereits abzeichnet, unterscheiden sich die Grenzen der verschiedene Alignments mitunter erheblich voneinander. So erstrecken sich die 429 Alignments über insgesamt 225 unterschiedliche Sequenzbereiche und umfassen 98 unterschiedliche N- sowie 111 verschiedene C-Termini. Als mögliche Ursache der Varianz kommen verschiedene Faktoren in Betracht:

- Die Domänen eines Multidomänenproteins grenzen nur selten unmittelbar aneinander, sondern sind meist über sogenannte Interdomänen-Linker miteinander verknüpft [76]. Die Ähnlichkeit zwischen den Domänen verschiedener Proteine kann sich bis in diese Sequenzbereiche hinein erstrecken. Beziehen die verschiedenen Alignments zu einer Domäne unterschiedliche Anteile des Interdomänen-Linkers mit ein, sind variierende Termini die Folge.
- Größere Variationen der Alignmentregionen sind auf Insertions- bzw. Deletionsereignisse zurückzuführen, die eine entsprechend verlängerte bzw. verkürzte Variante der jeweiligen Domänen zur Folge haben. Wo Strukturdaten verfügbar sind, zeigt sich, daß die dreidimensionale Struktur der betroffenen Domäne in der Regel trotz Insertion erhalten bleibt. Die eingefügten Sequenzabschnitte befinden sich in der Regel in einer der Proteinoberfläche zugewandten Schleife (*Loop*) und haben daher nur wenig Einfluß auf die Struktur der eigentlichen Domäne. Dies erklärt auch, weshalb die eingefügten Sequenzabschnitte funktionell toleriert werden können [77].

- Eine besondere Form der Insertion sind ineinander verschachtelte Domänen (*nested domains*). Manche Domänen sind durch die Insertion einer oder mehrerer anderer Domänen unterbrochen. So liegt beispielsweise die IMP-Dehydrogenase/GMP-Reduktase-Domäne (IMPDH) in den meisten Sequenzen, die diese Domäne enthalten, kontinuierlich vor. In einigen Fällen ist die lineare Sequenz der IMPDH-Domäne jedoch durch die Insertion von Cystathionin-β-Synthase-Domänen (CBS) unterbrochen [77] (vgl. Abbildung 3-5). Solche diskontinuierlichen Domänen können zu partiellen Alignments führen, die nur einen Teil der Domäne umfassen.

Abbildung 3-5: Manche Domänen liegen aufgrund der Insertion einer oder mehrerer anderer Domänen diskontinuierlich vor. So ist beispielsweise die IMP-Dehydrogenase/GMP-Reduktase-Domäne (IMPDH) in einigen Fällen durch die Insertion von Cystathionin-β-Synthase-Domänen (CBS) unterbrochen.

Quelle: Sanger Institute (www.sanger.ac.uk/Software/Pfam/)

Um unter der Vielzahl verschiedener Termini die Domänengrenzen zu ermitteln, wurde versucht, den unterschiedlichen Ursachen der Varianz Rechnung zu tragen. Das dazu entwickelte Verfahren gliedert sich in zwei aufeinanderfolgende Schritte. Während des ersten Schrittes werden zunächst nur die Alignment-Termini betrachtet, unabhängig davon, über welche Sequenzbereiche sich die dazugehörigen Alignments erstrecken. Auf diese Weise ist es möglich, auch die Termini sehr unterschiedlich langer Alignments miteinander zu korrelieren. Ziel des Schrittes ist es, alle geringfügigen Variationen eines Terminus, wie sie beispielsweise infolge von Interdomänen-Linker entstehen, auf eine einzelne, repräsentierende Position zurückzuführen.

Eine Analyse der Interdomänen-Linker ergab, daß ihre durchschnittliche Länge $10 \pm 5{,}8$ Aminosäuren beträgt [78]. Es wird daher davon ausgegangen, daß alle Alignment-Termini, die sich innerhalb dieser Reichweite zueinander befinden, verschiedene Varianten desselben Domänenterminus repräsentieren. Dementsprechend wird die Sequenz in Abschnitte von je 15 Aminosäuren unterteilt. Anschließend werden alle Termini eines jeweiligen Typs, die sich in-

nerhalb des gleichen Abschnitts befinden, durch einen einzelnen Repräsentanten ersetzt (vgl. Abschnitt 2.5). Als Repräsentant wird der Terminus innerhalb des Abschnitts gewählt, dessen Position am häufigsten als N- bzw. C-Terminus eines Alignments fungiert, da dieser eine erhöhte Wahrscheinlichkeit besitzt, die Grenze der Domäne zu markieren. Abbildung 3-6 sowie Tabelle 3-5 veranschaulichen, wie sich die Zahl der Alignment-Termini infolgedessen verringert.

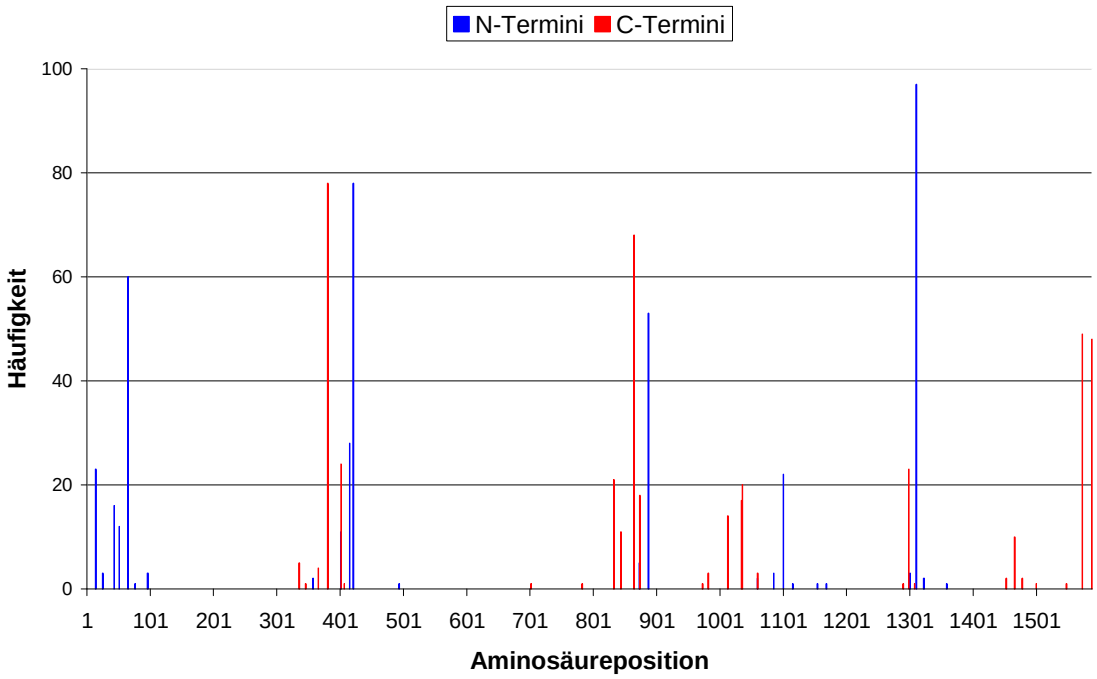


Abbildung 3-6: Graphische Repräsentation der N- (blau) und C-Termini (rot) aller nach dem ersten Schritt der Domänenstrukturvorhersage zur Swiss-Prot Sequenz ARO1_YEAST verbliebenen Alignmentregionen.

Im zweiten Schritt werden nicht mehr nur die Termini sondern auch Lage und Ausdehnung der einzelnen Alignments betrachtet. Ziel dieses Schrittes ist es, auch größere Variationen einer Alignmentregion, wie sie infolge von Insertionen entstehen, auf einen einzelnen, repräsentativen Bereich der Sequenz zurückzuführen. Dazu werden alle nach dem ersten Schritt verbliebenen Alignmentregionen miteinander verglichen und entsprechend ihrer Ähnlichkeit gruppiert. Die Ähnlichkeit der Alignmentregionen bemißt sich an der Anzahl ihrer nicht überlappenden Aminosäuren. Je geringer diese ausfällt, desto ähnlicher sind die Sequenzbereiche, über die sich die Alignments erstrecken. Um bloße Längenvarianten eines Sequenzbereichs von solchen Alignments zu un-

terscheiden, die auch benachbarte Domänen mit einschließen, wird eine maximal zulässige Längendifferenz definiert. Diese orientiert sich erneut an der kürzesten im Datensatz definierten Domäne (vgl. Abschnitt 3.3.3.1). Die SH3-Domäne umfaßt etwa 60 Aminosäuren [67]. Um während der Clusterung keine benachbarten Domänen in das Cluster zu inkorporieren, muß der zulässige Längenunterschied folglich unterhalb dieses Wertes rangieren. Nach Tests mit verschiedenen Grenzwerten wurde die N- wie C-terminale Differenz letztlich auf 45 Aminosäuren beschränkt und entspricht somit der Größe dreier Abschnitte des vorangegangen Schrittes.

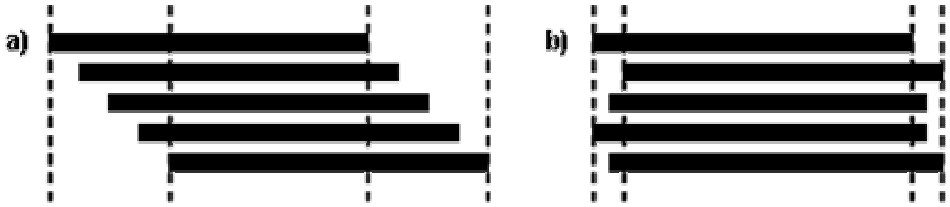


Abbildung 3-7: a) Beim *single linkage* Verfahren reicht es aus, wenn die beiden ähnlichsten Objekte zweier Cluster dem Grenzwert genügen. Dies kann im Verlauf der Clusterung zur Kettenbildung führen, in deren Folge auch benachbarte Domänen mit eingeschlossen werden. b) Das *complete linkage* Verfahren berücksichtigt hingegen alle Einzelbeziehungen. Homologe Cluster sind die Folge.

Zur Clusterung der ähnlichen Alignmentabschnitte wird das *complete linkage* Verfahren verwendet, da dieses im Gegensatz zum *single linkage* Verfahren, das im späteren Verlauf Anwendung findet, nicht zur Kettenbildung tendiert (vgl. Abschnitt 3.5 und Abbildung 3-7a). So wird sichergestellt, daß die gruppierten Alignmentregionen tatsächlich dieselbe Sequenzregion umfassen (vgl. Abbildung 3-7b). Nach Abschluß der Clusterung wird für jedes Cluster der N- und C-Terminus mit der jeweils größten Häufigkeit innerhalb des Clusters ermittelt. Diese bilden fortan die Grenzen des im Cluster enthaltenen Sequenzbereichs. Da infolge des Prozesses Sequenzabschnitte entstehen können, deren Längendifferenz nun innerhalb des Grenzwertes von 45 Aminosäuren rangiert, wird die Clusterung anschließend nochmals mit den so erhaltenen Regionen wiederholt. Dies erfolgt solange, bis keine Änderung der Grenzen mehr eintritt. Abbildung 3-8 zeigt das Endergebnis dieses Prozesses.

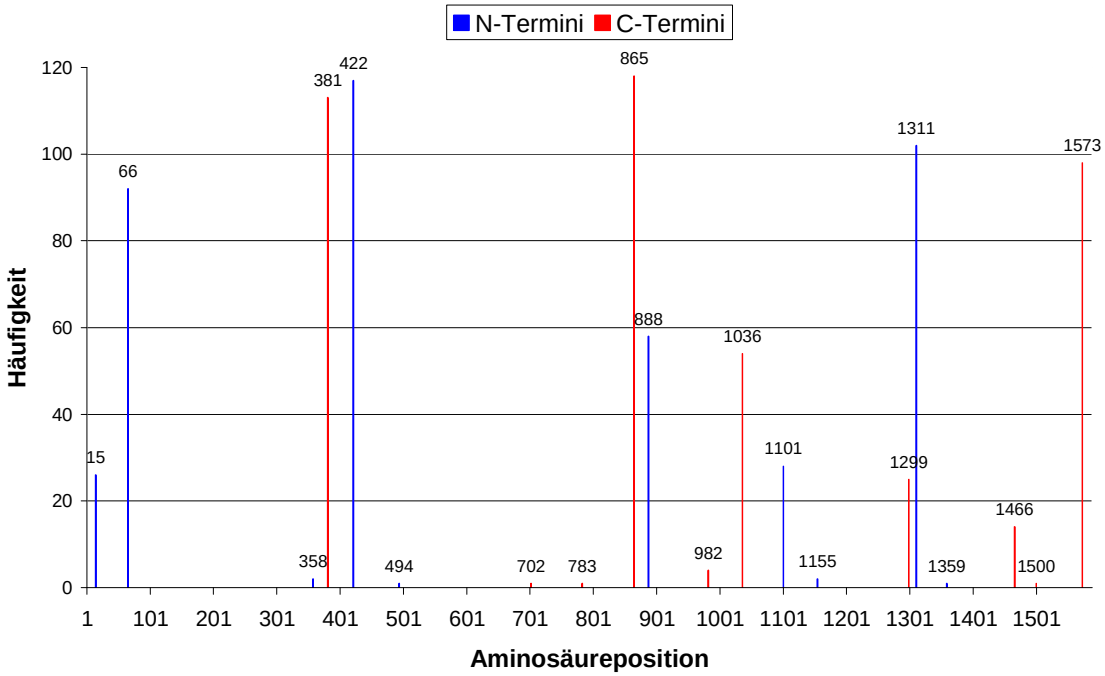


Abbildung 3-8: Graphische Repräsentation der N- (blau) und C-Termini (rot) aller nach dem zweiten Schritt der Domänenstrukturvorhersage zur Swiss-Prot Sequenz ARO1_YEAST verbliebenen Alignmentregionen.

Die graphische Darstellung veranschaulicht, daß sich die Zahl der Alignment-Termini infolge des letzten Schrittes nochmals deutlich reduziert hat. Durch die Kombination beider Schritte lassen sich die ursprünglich 225 Alignmentregionen auf nur 17 unterschiedliche Sequenzbereiche zurückführen (vgl. Tabelle 3-5). Tabelle 3-6 enthält eine Auflistung der verbliebenen Alignmentregionen sowie die Zahl der Alignments, die sich über jede dieser Regionen erstreckt.

Tabelle 3-5: Reduktion der Alignmentregionen im Verlauf der Methode

Stufe der Domänen-strukturvorhersage	Anzahl unterschiedlicher Alignmentregionen	
	ARO1_YEAST	gesamter Datensatz
Ausgangswert	225	4.120.028
1. Schritt	62	1.169.505
2. Schritt	17	306.042

Tabelle 3-6: Durch die Kombination beider Methoden läßt sich die Zahl der unterschiedlichen Alignmentregionen von ursprünglich 225 auf 17 reduzieren. Die Tabelle zeigt deren Lage sowie die Zahl der Alignments, die sich über jede der Regionen erstreckt. Alignmentregionen, die sich über ähnliche Sequenzbereiche erstrecken, sind zu Gruppen zusammengefaßt. Die Region innerhalb jeder Gruppe, welche die Mehrzahl der Alignments auf sich vereint, ist durch Fettdruck hervorgehoben.

Gruppe	Alignmentregion		Zahl der Alignments
	N-Terminus	C-Terminus	
1	15	1573	5
2	15	381	21
	66	**381**	**92**
3	358	865	2
	422	702	1
	422	783	1
	422	**865**	**115**
	494	865	1
4	888	982	4
	888	**1036**	**54**
5	**1101**	**1299**	**25**
6	1101	1500	1
	1101	1573	2
	1155	1573	2
7	1311	1466	14
	1311	**1573**	**88**
	1359	1573	1

Die 17 verbliebenen Alignmentregionen lassen sich entsprechend des von ihnen erfaßten Sequenzbereichs in sieben Gruppen gliedern (vgl. Tabelle 3-6). Zwei dieser Gruppen (Gruppe 1 und 6) repräsentieren Alignments, die sich über einen Bereich von mehreren Sequenzdomänen erstrecken. Die übrigen fünf Gruppen zeigen von einer Ausnahme abgesehen keine Überlappung zueinander und verkörpern die verbliebenen Längenvarianten zu den fünf Sequenzdomänen des Enzyms. Anhand der Alignmentzahl läßt sich für jede dieser Gruppen die Sequenzregion bestimmen, die am wahrscheinlichsten die Grenzen der jeweiligen Domäne repräsentiert. Tabelle 3-7 enthält eine Gegenüberstellung dieser Werte zu den Angaben aus anderen Datenbanken.

Tabelle 3-7: Gegenüberstellung der zu ARO1_YEAST ermittelten Domänengrenzen und den Angaben aus den Datenbanken Swiss-Prot, Pfam und PRODOM

Domäne	Datenbank			eigene
	Swiss-Prot	Pfam	Prodom	Methode
3-Dehydroquinat Synthase	1–392	17-367	23-399	66-381
3-Phosphoshikimat 1-Carboxyvinyltransferase	404–866	411-865	412-701 703-830 845-877	422-865
Shikimat Kinase	887–1060	897-1064	895-1032 1033-1101	888-1036
3-Dehydroquinat Dehydratase	1061–1293	1077-1297	1102-1302	1101-1299
Shikimat Dehydrogenase	1306-1588	1324-1574	1311-1394 1403-1502 1531-1584	1311-1573

Alle übrigen der verbliebenen Alignmentregionen repräsentieren vergleichsweise selten auftretende Varianten der jeweiligen Domäne, deren Längenunterschiede mehr als 45 Aminosäuren betragen. Bei einigen von diesen liegt die Differenz nur geringfügig oberhalb dieses Grenzwertes (1311-1573/1359-1573). Es sind jedoch auch Alignments darunter, deren Längenunterschied mehr als 100 Aminosäuren beträgt (422-702/422-865, 1311-1466/1311-1573).

Die biologische Bedeutung dieser im Vergleich zur durchschnittlichen Länge soviel kürzeren Domänenvarianten ist unklar. Es muß jedoch zunächst davon ausgegangen werden, daß auch sie funktionsfähige Formen der jeweiligen Domänen repräsentieren. Wenn diese bereits zur Ausübung der Funktion ausreichen, stellt sich die Frage, worin sich die längeren Formen der Domänen von diesen unterscheiden. Handelt es sich bei den zusätzlichen Sequenzbereichen lediglich um funktionsneutrale Insertionen, oder beeinflußt sie möglicherweise die Art der Reaktion? Manche dem Kernbereich einer katalytischen Domäne angehängten Sequenzabschnitte scheinen für eine bestimmte EC-Klasse charakteristisch zu sein und kommen nur in dieser vor. Die Längenvarianten könnten somit auf die mögliche Existenz submodularer Funktionseinheiten hinweisen, die dem eigentlich katalytischen Abschnitt der Domäne vor- oder nachgelagert sind und so beispielsweise die Spezifität der Reaktion beeinflussen.

Ein möglicher Hinweis auf die Existenz solcher submodularen Einheiten ergibt sich aus dem Vergleich zu den von Prodom ermittelten Domänen (vgl. Tabelle 3-7). Während Swiss-Prot und Pfam je fünf Domänen angeben, deren Grenzen der jeweils häufigsten Alignmentvariante der eigenen Methode entsprechen, unterteilt Prodom die Sequenz in zehn separate Domänen. Zwei von diesen korrelieren mit den Angaben der anderen Datenbanken. Bei den übrigen Sequenzbereichen zeigt Prodom jedoch eine feinere Gliederung, welche die jeweilige Region nochmals in zwei bis drei separate Domänen unterteilt. Zumindest in einem Fall besteht eine auffallende Übereinstimmung zwischen einer von Prodom postulierten Domäne (412-701) und einer der eigenen Alignmentvarianten zu diesem Sequenzbereich (422-702) (vgl. Tabelle 3-6).

In Anbetracht der zum Teil erheblichen Längenunterschiede und den möglicherweise damit verbundenen Unterschieden in der Reaktion werden daher nicht nur die fünf häufigsten Formen der Domänen, sondern alle zu diesem Zeitpunkt verbliebenen Alignmentvarianten der Clusteranalyse unterzogen. Für den gesamten Datensatz ergeben sich so 306.042 zu gruppierende Sequenzabschnitte (vgl. Tabelle 3-5). Die zu deren Ermittlung benötigte Rechenzeit betrug weniger als 3 ½ Stunden (Angabe bezieht sich auf eine SunFire V880 unter Verwendung eines der insgesamt sechs UltraSparc III/750 Prozessoren).

3.5 Die Clusteranalyse

Nachdem im vorangegangenen Schritt die Sequenzbereiche bestimmt wurden, über die ein jeweiliges Enzym Ähnlichkeit zu anderen Enzymen des Datensatzes besitzt, geht es im folgenden darum, die analogen Sequenzabschnitte der verschiedenen Enzyme entsprechend ihrer Ähnlichkeit zu gruppieren.

Zur Clusterung der korrespondierenden Sequenzabschnitte wird das *single linkage* Verfahren verwendet. Bei diesem Verfahren werden jeweils diejenigen zwei Cluster vereint, deren ähnlichste Nachbarobjekte die geringste Distanz zueinander aufweisen, d. h. das Verhältnis zwischen je einem Objekt des einen und des anderen Clusters definiert den Abstand zwischen beiden Clustern. Die übrigen Objekte der Cluster spielen dagegen keine Rolle (vgl. Abschnitt 2.6.2.2).

Dadurch, daß nur zwei nahe beieinander liegende Einzelobjekte über die Fusionierung zweier Cluster entscheiden, birgt diese Methode die Gefahr der Bildung langgestreckter Cluster (Kettenbildung). Es kann jedoch dem

angestrebten Ziel der gewünschten Klassifizierung widersprechen, wenn zwei Objekte ein- und desselben Clusters eine größere Distanz zueinander aufweisen als Objekte verschiedener Cluster. Der Vorteil dieser Methode besteht indes darin, daß sie besonders geeignet ist, um auch entfernt verwandte Objekte zu identifizieren. Die Eigenschaft, Objekte auch ohne direkte Ähnlichkeit über Intermediärobjekte miteinander assoziieren zu können, entspricht dem Prinzip der Transitivität und ermöglicht es so, Homologe über ein Maß der direkten Sequenzidentität hinaus zu ermitteln.

3.5.1 Übersicht über den Verlauf der Clusterung

Der Clusterprozeß wird mit der Sequenzpaarung begonnen, welche die größte Ähnlichkeit zueinander aufweist, wobei der beim ungefilterten „alle vs. alle"-Alignment ermittelte *E-value* verwendet wird, um den Grad der Identität zu quantifizieren. Das dazugehörige lokale Alignment gibt Auskunft darüber, welche Sequenzbereiche beider Enzyme einander entsprechen. Diese werden daraufhin zu einem Cluster vereint. Im Anschluß daran wird der Signifikanzlevel schrittweise erniedrigt, und es werden nach und nach weitere Ähnlichkeiten berücksichtigt. Um den Verlauf der Clusterung verfolgen zu können, werden Zwischenergebnisse für insgesamt zehn verschiedene Grenzwerte gespeichert.

In Tabelle 3-8 ist eine Zusammenfassung des Clusterungsverlaufs dargestellt. Für jeden der zehn verwendeten Grenzwerte ist die Gesamtzahl der bis dahin erhaltenen Gruppierungen angegeben. Darüber hinaus ist dargestellt, wie viele Cluster einer jeweiligen Größe entstehen. Abbildung 3-9 veranschaulicht, wie sich die Zahl der Cluster mit sinkender Signifikanz des zur Clusterung verwendeten Grenzwertes sukzessive verringert. Bei niedrigen *E-values* werden zunächst nur solche Sequenzabschnitte in einem gemeinsamen Cluster vereint, die eine sehr hohe Sequenzübereinstimmung besitzen. Aus biologischer Sicht ergeben sich so Gruppen hochkonservierter Unterfamilien. Mit abnehmender Signifikanz des Grenzwertes werden schrittweise auch geringere Ähnlichkeiten berücksichtigt, so daß einzelne Cluster zu größeren Gruppen zusammengefaßt werden und sich ihre Zahl infolgedessen verringert. So verteilen sich die

Tabelle 3-8: Anzahl und Größen der Cluster bei verschiedenen Grenzwerten

Grenzwert	Clustergröße (Objektzahl)							Gesamtzahl der Cluster
	> 100	51-100	21-50	11-20	6-10	2-5	1	
10^{-180}	24	44	136	240	576	4.897	272.297	278.214
10^{-110}	73	115	280	460	868	5.908	241.898	249.602
10^{-80}	116	151	363	514	950	5.615	224.095	231.804
10^{-50}	198	211	445	563	903	4.685	197.103	204.108
10^{-35}	258	246	485	546	823	4.203	174.193	180.754
10^{-25}	296	265	488	533	840	4.167	149.773	156.362
10^{-15}	340	271	500	567	822	4.464	112.163	119.127
10^{-10}	376	285	481	578	877	4.745	82.013	89.355
10^{-5}	404	270	488	563	918	5.496	37.972	46.111
10^{-2}	370	248	457	553	975	6.283	2.745	11.631

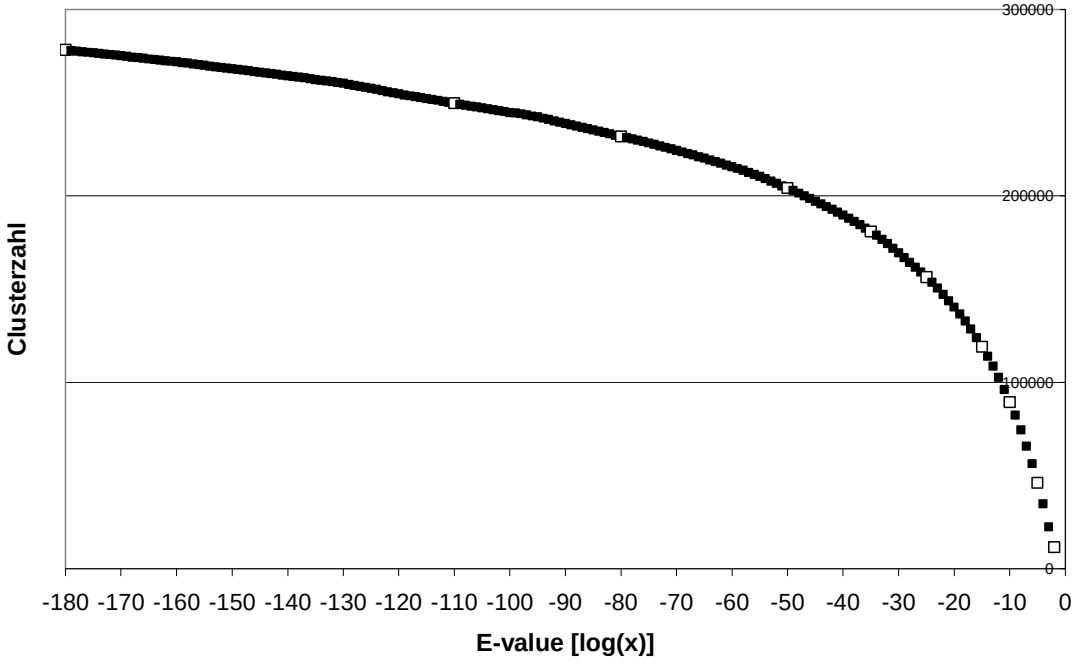


Abbildung 3-9: Entwicklung der Clusterzahl in Abhängigkeit vom *E-value*. Zu Beginn werden nur sehr ähnliche Sequenzabschnitte gruppiert. Mit sinkender Signifikanz des Grenzwertes werden dann auch geringere Ähnlichkeiten berücksichtigt, so daß einzelne Cluster zu größeren Gruppen zusammengefaßt werden. Die weißen Datenpunkte (□) markieren die Grenzwerte, für die ein Zwischenergebnis gespeichert wurde.

306.042 zu gruppierenden Sequenzabschnitte bei einem Grenzwert von 10^{-180} noch auf 278.214 separate Cluster, während bei einem Wert von 10^{-2} lediglich 11.631 Cluster verbleiben. Im gleichen Maße, mit dem sich die Gesamtzahl der Cluster verringert, nimmt die Größe der entstandenen Cluster zu. So erhöht sich die Anzahl der Cluster mit mehr als hundert Objekten kontinuierlich von 24 (10^{-180}) auf über 400 bei einem Grenzwert von 10^{-5}. Bei einem Wert von 10^{-2} verringert sich deren Zahl jedoch wieder, was darauf hinweist, daß bei Grenzwerten oberhalb von 10^{-5} mehrere große Cluster miteinander fusioniert werden (vgl. auch Abbildung 3-10).

3.5.2 Bestimmung eines Grenzwertes

Das Grundprinzip der Clusteranalyse besteht darin, die einzelnen Objekte eines Datensatzes derart in Gruppen (Cluster) zu sortieren, daß gruppeninterne Ähnlichkeiten maximiert werden, während die Ähnlichkeiten zwischen den verschiedenen Clustern minimiert werden. Dabei ergeben sich sowohl die Zuordnungen der einzelnen Objekte als auch die Anzahl der erhaltenen Cluster allein anhand des verwendeten Datenmaterials.

Der entscheidende Schritt im Verlauf der Analyse ist es, diejenige Stufe festzulegen, nach welcher die Prozedur abzubrechen ist. Man hofft, daß sich zu einem gegebenen Zeitpunkt eine intuitive, d. h. natürliche Gruppierung derart ergibt, daß man eine bearbeitungsfähige Anzahl verschiedener Cluster erhält. Ebenso wie es mehrere verschiedene Möglichkeiten zur Bestimmung der Distanz zwischen den Clustern gibt, existieren auch unterschiedliche Ansätze zur Feststellung der finalen Stufe einer Clusteranalyse, wobei die verschiedenen Methoden nicht notwendigerweise zum selben Ergebnis führen müssen. Derartige Orientierungsregeln sind aber notwendig, da die „richtige“ Anzahl von Clustern nur selten offensichtlich ist.

Eine mögliche Methode, die Abbruchstufe festzulegen, ist eine graphischen Darstellung der Clusterzahl als Funktion der Analysestufe. Sollte sich zu irgendeinem Zeitpunkt der Clusterung die Gesamtzahl der verbleibenden Cluster abrupt verringern, so kann dieser „Kollaps“ des Clusterbaums als Indiz dafür gewertet werden, daß von diesem Augenblick an auch Sequenzanschnitte miteinander gruppiert werden, die das Kriterium der Homologie nicht erfüllen.

Der Verlauf des entsprechenden Graphen (vgl. Abbildung 3-9) zeigt zwar, daß sich die Abnahme der Clusterzahl zu den höheren Grenzwerten hin beschleunigt, dabei handelt es sich jedoch um einen kontinuierlichen Prozeß, der keine explizite Abbruchstufe erkennen läßt. Ursache für die beschleunigte Abnahme der Clusterzahl ist die größere Zahl der Alignments, die zu den höheren Grenzwerten existiert (vgl. Abbildung 3-2). So besitzen ⅓ aller beim Sequenzvergleich erhaltenen Alignments *E-values* > 10^{-25}. Infolgedessen werden mit abnehmender Signifikanz des zur Clusterung verwendeten Grenzwertes auch mehr Sequenzabschnitte respektive Cluster miteinander assoziiert, als dies bei niedrigen *E-values* der Fall ist.

Eine weitere Möglichkeit, eine Abbruchstufe festzulegen, besteht darin, die Größenentwicklung des jeweils größten Clusters zu verfolgen. Dazu wird nach jeder Analysestufe die Größe des jeweils objektreichsten Clusters ermittelt (vgl. Abbildung 3-10).

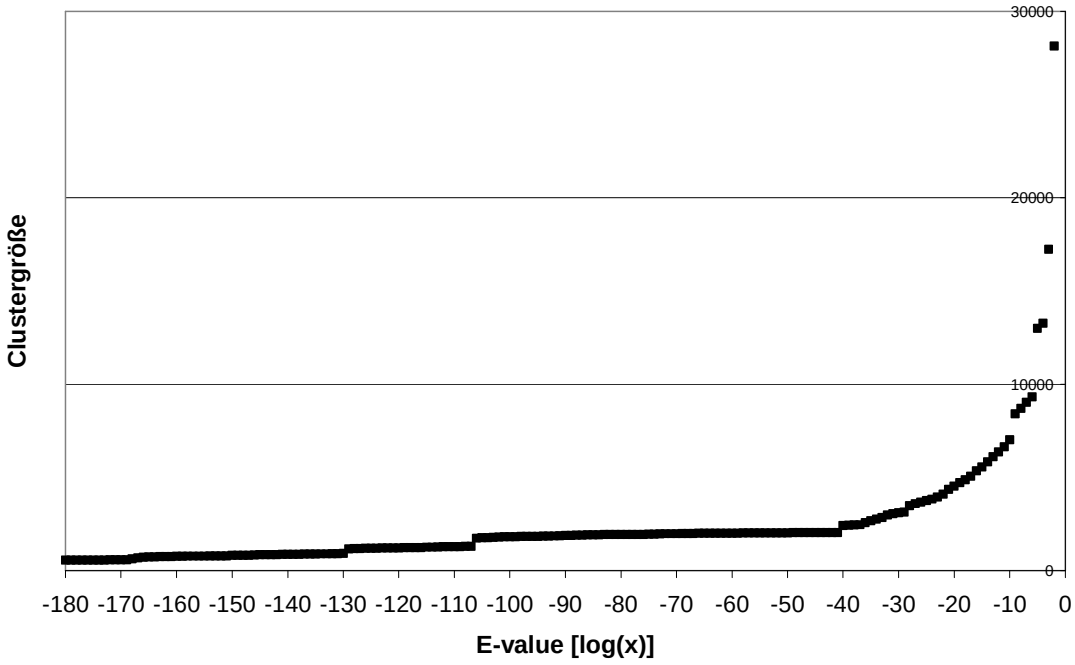


Abbildung 3-10: Graphische Repräsentation der Größenentwicklung des jeweils größten Clusters in Abhängigkeit vom *E-value*. Bei niedrigen *E-values* ist nur ein geringer Größenzuwachs zu verzeichnen. Bei höheren Grenzwerten nimmt die Clustergröße hingegen sprunghaft zu.

Zu Beginn der Analyse werden nur kleine, sehr ähnliche Cluster vereinigt, so daß die Größenänderung nur gering ausfällt. Am Ende der Prozedur sind hingegen nur noch wenige Cluster vorhanden, die durch größere Abstände voneinander getrennt sind. Ist eine Stufe auszumachen, bei welcher die Clustergröße einen markanten Sprung macht, so ist die Analyse unmittelbar vor dieser Stufe zu beenden. Dabei kommt es auch vor, daß mehrere Stellen als Abbruchpunkt in Frage kommen, was die Subjektivität der Methode ausmacht. So kommen anhand der Abbildung 3-10 theoretisch mindestens drei Stellen in Betracht: nach einem *E-value* von 10^{-29},10^{-10} und 10^{-6}. Den Ausschlag dafür, daß letztlich ein *E-value* von 10^{-10} als Ausgangswert für die Betrachtung der Cluster gewählt wurde, gab einerseits die Anzahl der resultierenden Cluster, die ansonsten unerwünscht groß ausgefallen wäre und eine Auswertung zusätzlich erschwert hätte, und andererseits die Verläßlichkeit der erhaltenen Gruppierungen, die bei niedrigen *E-values* generell höher ist.

3.6 Ergebnis der Clusteranalyse

3.6.1 Zahl der erhaltenen EC-Kombinationen

Anhand des Clusterergebnisses soll untersucht werden, ob Enzyme, die homolog zueinander sind, gemäß EC-Klassifikation jedoch unterschiedliche Reaktionen katalysieren, ähnliche enzymatische Reaktionsmechanismen zu deren Realisierung verwenden.

Tabelle 3-9 enthält eine Übersicht darüber, wie viele Paarungen zwischen verschiedenen EC-Klassen im Verlauf der Clusteranalyse zustande kommen, wie viele unterschiedliche EC-Klassen daran beteiligt sind und auf welcher Ebene der EC-Klassifikation sich diese voneinander unterscheiden. Da es nicht möglich ist, die EC-Paarungen, die sich auf katalytische Domänen beziehen, automatisiert von solchen zu unterscheiden, die auf anderen Enzymdomänen beruhen, schließen die angegebenen Werte alle Gruppierungen unterschiedlicher EC-Klassen mit ein, unabhängig davon, über welche Domänen sich diese ergeben. Die Werte vermitteln dennoch einen Eindruck davon, in welchem Umfang Homologien zwischen den Sequenzen unterschiedlicher EC-Klassen bestehen.

Tabelle 3-9: Die Tabelle enthält eine Übersicht darüber, wie viele Paarungen zwischen verschiedenen EC-Klassen bei den unterschiedlichen Grenzwerten entstehen, wie viele EC-Klassen daran beteiligt sind und auf welcher Ebene der EC-Klassifikation sich diese voneinander unterscheiden.

Grenzwert	Zahl der Verbindungen für jede Ebene der EC-Klassifizierung				Zahl der EC-Klassen
	Hauptklasse	Subklasse	Sub-Subklasse	Sub-Sub-Subklasse	
10^{-180}	17	31	60	192	268
10^{-110}	52	105	154	557	570
10^{-80}	76	183	260	960	742
10^{-50}	177	387	514	1.938	1024
10^{-35}	388	836	1088	3.729	1181
10^{-25}	511	1.211	1.547	4.738	1313
10^{-15}	863	2.031	2.470	6.268	1456
10^{-10}	1.283	2.880	3.420	7.614	1586
10^{-5}	2.513	5.114	5.933	10.664	1712
10^{-2}	58.608	78.369	80.492	89.203	1816

Es zeigt sich, daß bereits beim niedrigsten der zehn verwendeten Grenzwerte Sequenzabschnitte miteinander gruppiert werden, die Enzymen unterschiedlicher EC-Klassifikation angehören. So bestehen bei einem Grenzwert von 10^{-180} bereits Verbindungen zwischen Sequenzen aus 268 verschiedenen EC-Klassen. Dies entspricht 12,5% der insgesamt betrachteten Enzymklassen. Mit steigendem Grenzwert erhöht sich deren Zahl noch, so daß bei einem Wert von 10^{-50} schon fast die Hälfte aller im Datensatz enthaltenen EC-Klassen in Kombination miteinander auftreten, während beim höchsten Grenzwert von 10^{-2} sogar 85% aller EC-Klassen Ähnlichkeiten zu den Sequenzen anderen Enzymklassen zeigen. Somit verbleiben letztlich nur 327 EC-Klassen im Datensatz, deren Sequenzen auch bei Verwendung des höchsten Grenzwertes keine Ähnlichkeit zu Sequenzen anderer EC-Klassen aufweisen.

Die Mehrzahl der EC-Kombinationen, die sich während der Clusterung ergeben, unterscheidet sich lediglich in der letzten Ziffer ihrer EC-Klassifikation. So differieren 132 der insgesamt 192 bei 10^{-180} vorkommenden EC-Paarungen auf

der untersten Ebene der EC-Klassifikation, 175 gehören zumindest derselben Hauptklasse an und lediglich 17 der gemeinsam auftretenden EC-Klassen unterscheiden sich bereits in ihrer Hauptklassenzuordnung. Mit steigendem Grenzwert erhöhen sich auch diese Werte, doch das Verhältnis zwischen ihnen bleibt weitgehend konstant. Erst bei einem Grenzwert von 10^{-2} ist ein markanter Sprung zu verzeichnen. So steigt die Zahl der EC-Kombinationen bei diesem Wert auf allen vier Ebenen der EC-Klassifikation sprunghaft an, während die Zahl der daran beteiligten EC-Klassen nur eine moderate Steigerung erfährt. Gleichzeitig kehren sich die Verhältnisse um, so daß bei diesem Grenzwert erstmals mehr EC-Kombinationen bestehen, die in der Haupt- bzw. Subklasse differieren, als solche, die sich auf den beiden unteren Ebenen der EC-Klassifikation unterscheiden. Diese Entwicklung ist ebenfalls auf die zuvor beschriebene Fusionierung mehrerer, sehr großer Cluster zurückzuführen, die bei diesem Grenzwert zu beobachten ist (vgl. Abschnitt 3.5.1). Eine detaillierte Darstellung der Clusterergebnisse für alle zehn verwendeten Grenzwerte ist der beigefügten CD zu entnehmen.

3.6.2 Art der erhaltenen EC-Kombinationen

Die Häufigkeit, mit der die einzelnen Haupt-, Sub- und Sub-Sub-Klassen miteinander gruppiert werden, läßt erste Rückschlüsse auf die Verwandtschafsbeziehungen zwischen den Klassen und den von ihnen kodierten Reaktionen zu. Einige EC-Klassen zeigen ein sehr hohes Maß der Korrelation und treten entsprechend häufig in gemeinsamen Clustern auf. Andere EC-Kombinationen sind dagegen selten oder kommen gar nicht vor. Die folgenden Tabellen geben eine Übersicht darüber, welche Enzymklassen überdurchschnittlich häufig miteinander gruppiert werden und infolgedessen eine Verwandtschaft ihrer Funktionen vermuten lassen.

Tabelle 3-10 enthält eine Auflistung aller EC-Hauptklassenpaarungen. Neben der Anzahl der Enzymklassen, die zu jeder der Hauptklassen verfügbar ist und aus denen sich die Zahl theoretisch möglicher EC-Kombinationen ableitet, ist für fünf Grenzwerte die Zahl der tatsächlichen erhaltenen Paarungen sowie die daraus resultierende Wahrscheinlichkeit für die jeweiligen EC-Kombination angegeben. Als Vergleichswert dient die mittlere Wahrscheinlichkeit für das Zustandekommen einer EC-Paarung beim jeweiligen Grenzwert.

Tabelle 3-10: Übersicht über die Häufigkeit der verschiedenen EC-Hauptklassenpaarung. Zu jeder EC-Kombination sind die Zahl der erhaltenen Paarungen sowie die daraus resultierende Wahrscheinlichkeit für die jeweilige Kombination angegeben.

EC-Paarung		Zahl der EC-Klassen		Zahl der erhaltenen EC-Paarungen bei verschiedenen Grenzwerten					Wahrscheinlichkeit einer Paarung bei verschiedenen Grenzwerten				
EC 1	EC 2	EC 1	EC 2	10^{-110}	10^{-35}	10^{-10}	10^{-5}	10^{-2}	10^{-110}	10^{-35}	10^{-10}	10^{-5}	10^{-2}
1	**1**	532	532	694	1932	3514	4091	19259	0,005	0,014	0,025	0,029	0,136
1	**2**	532	585	5	23	194	344	14978	$2 * 10^{-5}$	$7 * 10^{-5}$	$6,2 * 10^{-4}$	0,001	0,048
1	**3**	532	649	2	12	48	103	12248	10^{-5}	$3 * 10^{-5}$	$1,4 * 10^{-4}$	$3 * 10^{-4}$	0,035
1	**4**	532	193	3	37	115	178	6626	$3 * 10^{-5}$	$3,6 * 10^{-4}$	0,001	0,002	0,065
1	**5**	532	91	4	33	120	150	2400	$8 * 10^{-5}$	$6,8 * 10^{-4}$	0,002	0,003	0,050
1	**6**	532	93	-	24	43	58	3240	-	$4,9 * 10^{-4}$	$8,7 * 10^{-4}$	0,001	0,065
2	**2**	585	585	673	968	1395	1822	5905	0,004	0,006	0,008	0,011	0,034
2	**3**	585	649	10	78	259	614	5828	$3 * 10^{-5}$	$2,1 * 10^{-4}$	$6,8 * 10^{-4}$	0,002	0,015
2	**4**	585	193	10	50	167	308	3835	$9 * 10^{-5}$	$4,4 * 10^{-4}$	0,001	0,003	0,034
2	**5**	585	91	2	26	46	103	1282	$4 * 10^{-5}$	$4,9 * 10^{-4}$	$8,6 * 10^{-4}$	0,002	0,024
2	**6**	585	93	3	16	60	129	1881	$6 * 10^{-5}$	$2,9 * 10^{-4}$	0,001	0,002	0,035
3	**3**	649	649	835	2004	2811	3492	5725	0,004	0,01	0,013	0,017	0,027
3	**4**	649	193	3	23	70	146	2450	$2 * 10^{-5}$	$1,8 * 10^{-4}$	$5,6 * 10^{-4}$	0,001	0,020
3	**5**	649	91	-	10	22	112	905	-	$1,7 * 10^{-4}$	$3,7 * 10^{-4}$	0,002	0,015
3	**6**	649	93	5	6	37	102	1226	$8 * 10^{-5}$	10^{-4}	$6,1 * 10^{-4}$	0,002	0,020
4	**4**	193	193	224	279	356	394	1090	0,012	0,015	0,019	0,021	0,058
4	**5**	193	91	-	9	45	68	604	-	$5,1 * 10^{-4}$	0,003	0,004	0,034
4	**6**	193	93	2	5	20	56	785	$1,1 * 10^{-4}$	$2,8 * 10^{-4}$	0,001	0,003	0,044
5	**5**	91	91	103	112	130	141	219	0,025	0,027	0,031	0,034	0,052
5	**6**	91	93	3	36	37	42	319	$3,5 * 10^{-4}$	0,004	0,004	0,005	0,038
6	**6**	93	93	119	189	268	354	541	0,027	0,043	0,061	0,081	0,124
mittlere Wahrscheinlichkeit einer EC-Paarung									**0,001**	**0,003**	**0,004**	**0,006**	**0,040**

Wie aus den Werten ersichtlich ist, gehören die meisten der miteinander gruppierten EC-Klassen der jeweils selben Hauptklasse an und unterscheiden sich lediglich in den unteren Ebenen ihrer EC-Klassifikation. Gruppierungen verschiedener Hauptklassen sind dagegen vergleichsweise selten und treten erst bei höheren Grenzwerten vermehrt in Erscheinung.

Bis zu einem Grenzwert von 10^{-5} weisen die Hauptklassen der Isomerasen (5. Hauptklasse) und Ligasen (6. Hauptklasse) die größte interne Homologie auf. So liegt die Wahrscheinlichkeit, daß zwei verschiedene Ligasen miteinander gruppiert werden, bei diesem Grenzwert bei 8,1%, während die mittlere Wahrscheinlichkeit für die Gruppierung zweier unterschiedlicher EC-Klassen lediglich 0,6% beträgt. Bei einem Grenzwert von 10^{-2} ändern sich die Verhältnisse, so daß nun die Oxidoreduktasen (1. Hauptklasse) die größte Wahrscheinlichkeit einer internen Paarung besitzen. Die Wahrscheinlichkeit ihrer Gruppierung steigt von ursprünglich 2,9% bei 10^{-5} auf 13,6% bei 10^{-2} an und liegt somit noch über dem Wert der Ligasen (12,4%). Die Wahrscheinlichkeit für die Gruppierung zweier Isomerasen erfährt dagegen nur noch eine geringe Steigerung und liegt mit 5,2% geringfügig oberhalb der mittleren Wahrscheinlichkeit von 4%. Die geringste Tendenz zur Ausbildung interner Paarungen

zeigen die Transferasen (2. Hauptklasse) und Hydrolasen (3. Hauptklasse), deren Wahrscheinlichkeitswerte 3,4% bzw. 2,7% nicht übersteigen.

Betrachtet man die Gruppierungen zwischen verschiedenen Hauptklassen, so zeigt sich, daß die Hauptklassen, die zur Bildung interner Paarungen tendieren, auch überdurchschnittlich häufig mit EC-Klassen anderer Hauptklassen assoziiert sind. Besonders häufig sind dabei Gruppierungen von Oxidoreduktasen mit Lyasen (4. Hauptklasse) bzw. mit Ligasen. So liegt die Wahrscheinlichkeit für diese Paarungen bei einem Grenzwert von 10^{-2} bei jeweils 6,5%. Um die Ursache für diese und andere EC-Kombinationen zu eruieren, werden im folgenden die erhaltenen Sub-Klassenpaarungen betrachtet.

Tabelle 3-11: Übersicht über die EC-Subklassenpaarung, die beim Grenzwert von 10^{-2} die größte Wahrscheinlichkeit einer Gruppierung aufweisen und sich in Haupt- oder Subklassenzuordnung unterscheiden.

EC-Paarung		Zahl der EC-Klassen		Zahl der erhaltenen EC-Paarungen bei verschiedenen Grenzwerten					Wahrscheinlichkeit einer Paarung bei verschiedenen Grenzwerten				
EC 1	EC 2	EC 1	EC 2	10^{-110}	10^{-35}	10^{-10}	10^{-5}	10^{-2}	10^{-110}	10^{-35}	10^{-10}	10^{-5}	10^{-2}
1.1	1.18	148	4					243					0,410
1.1	3.13	148	1				7	81				0,047	0,547
1.1	5.99	148	2					162					0,547
1.1	6.4	148	4			4	4	324			0,007	0,007	0,547
1.18	1.4	4	24			6	6	51			0,063	0,063	0,531
1.18	1.6	4	12			15	16	22			0,313	0,333	0,458
1.18	2.2	4	5					9					0,450
1.18	3.13	4	1					3					0,750
1.18	5.99	4	2					6					0,750
1.18	6.4	4	4					12					0,750
1.4	2.2	24	5					53					0,442
1.4	3.13	24	1					17					0,708
1.4	5.99	24	2					34					0,708
1.4	6.4	24	4					68					0,708
1.5	3.13	35	1				1	16				0,029	0,457
1.5	5.99	35	2					32					0,457
1.5	6.4	35	4					64					0,457
1.6	3.13	12	1				1	6				0,083	0,500
1.6	5.99	12	2					12					0,500
1.6	6.4	12	4					24					0,500
2.1	3.13	87	1					38					0,437
2.1	5.99	87	2				1	76				0,006	0,437
2.1	6.4	87	4	3	3	4	8	160	0,009	0,009	0,011	0,023	0,460
2.2	3.13	5	1					3					0,600
2.2	5.99	5	2					6					0,600
2.2	6.4	5	4					12					0,600
2.6	2.9	29	1					23					0,793
2.6	4.4	29	11		3	22	44	140		0,009	0,069	0,138	0,439
2.9	4.4	1	11					6					0,545
3.13	4.6	1	5					2					0,400
3.13	5.2	1	5					2					0,400
3.13	5.99	1	2					2					1,000
3.13	6.4	1	4					4					1,000
4.3	6.4	14	4				4	24				0,071	0,429
4.5	5.2	1	5			2	2	2			0,400	0,400	0,400
4.6	5.99	5	2					4					0,400
4.6	6.4	5	4					8					0,400
5.2	5.99	5	2			1	1	4			0,100	0,100	0,400
5.2	6.4	5	4					8					0,400
5.99	6.4	2	4					8					1,000
mittlere Wahrscheinlichkeit einer EC-Paarung									**0,001**	**0,003**	**0,004**	**0,006**	**0,040**

Tabelle 3-11 enthält eine Übersicht über die vierzig EC-Subklassenpaarungen, die beim Grenzwert von 10^{-2} die größte Wahrscheinlichkeit einer Gruppierung aufweisen und sich zudem in ihrer Haupt- oder Subklassenzuordnung unterscheiden. Darüber hinaus ist dargestellt, welche Wahrscheinlichkeitswerte die Kombinationen bei anderen Grenzwerten erzielen. Eine vollständige Auflistung aller bei diesem und den übrigen neun Grenzwerten erhaltenen Subklassenpaarungen ist der beigefügten CD zu entnehmen.

Anhand der obigen Darstellung wird deutlich, daß viele EC-Paarungen, die bei 10^{-2} eine sehr hohe Gruppierungswahrscheinlichkeit erreichen, erst bei Grenzwerten oberhalb von 10^{-5} zustande kommen und somit auf relativ geringer Sequenzähnlichkeit beruhen. Dies gilt auch für die Subklassenpaarungen 3.13/5.99, 3.13/6.4 und 5.99/6.4, die erstmals bei 10^{-2} registriert werden, dann jedoch eine Gruppierungswahrscheinlichkeit von jeweils 100% erreichen. Diese drei Subklassen sind überdies noch an vielen weiteren EC-Paarungen beteiligt, von denen die meisten ebenfalls erst bei Werten über 10^{-5} zustande kommen. Daraus folgt, daß all diese Paarungen das Resultat eines einzigen Clusters sind, welches durch die Fusion mehrerer zuvor noch getrennter Cluster entsteht. Eine Überprüfung anhand des Clusterergebnisses bestätigt dies.

Wie zuvor bereits beschrieben, kommt es oberhalb von 10^{-5} zur Fusionierung gleich mehrerer objektreicher Cluster (vgl. Tabelle 3-8 im Abschnitt 3.5.1), in deren Folge sich die Zahl der erhaltenen EC-Kombinationen deutlich erhöht (vgl. Tabelle 3-9 im Abschnitt 3.6.1). Die meisten der in Tabelle 3-11 aufgeführten EC-Paarungen sind das Resultat dieses Prozesses. Sie sind Teil des mit 28.131 Sequenzabschnitten objektreichsten Clusters dieses Grenzwertes (vgl. Abbildung 3-10). Diese gehören 6.929 unterschiedlichen Sequenzen an, die allen sechs Hauptklassen entstammen. In Anbetracht ihrer funktionellen Vielfalt erscheint es zunächst unwahrscheinlich, daß eine evolutionäre Verwandtschaft zwischen all diesen Enzymen besteht, und auch die Grenzwertbestimmung zur Clusteranalyse ergab, daß der Clusterungsprozeß bereits vor einem Wert von 10^{-2} abzubrechen ist (vgl. Abschnitt 3.5.2). Da eine derart funktionsübergreifende Verwandtschaft jedoch von großem biologischem Interesse wäre, wurde dennoch versucht, eine mögliche Ursache für die Gruppierung der Enzyme zu ermitteln.

Die Auswertung des Clusters erweist sich als schwierig. Aufgrund der hohen Sequenzzahl und der nur geringen Sequenzähnlichkeit können im multiplen Sequenzalignment keine konservierten Sequenzpositionen bestimmt werden. Auch ein Vergleich der übrigen zur Verfügung stehenden Informationen erbringt keinen konkreten Hinweis auf die Ursache ihrer Gruppierung. Somit bleibt lediglich der manuelle Vergleich aller am Cluster beteiligten Enzymfunktionen. Dies ist aufgrund der Vielzahl der vertretenen Enzymklassen jedoch nicht zu realisieren. Um zumindest einen Anhaltspunkt zu erhalten, wurde die Betrachtung daher auf die drei zuvor genannten Subklassen 3.13, 5.99 und 6.4 beschränkt (vgl. Abschnitt 3.6.2). Tabelle 3-12 enthält eine Übersicht darüber, welche Enzyme durch die drei Subklassen repräsentiert werden und welche Reaktionen diese katalysieren.

Tabelle 3-12: Übersicht über die zur Betrachtung ausgewählte EC-Klassen und den von ihnen katalysierten Reaktionen

EC-Klasse	Enzym	katalysierte Reaktion
3.13.1.1	UDP-Sulfoquinovose Synthase	UDP-Glucose + Sulfit = UDP-6-Sulfoquinovose + H_2O
5.99.1.2	DNA Topoisomerase	ATP-unabhängige Spaltung von einstrangiger DNA, gefolgt von Passage und Wiederverknüpfung
5.99.1.3	DNA Topoisomerase (ATP-hydrolysierend)	ATP-abhängige Spaltung, Passage und Wiederverknüpfung von doppelstrangiger DNA
6.4.1.1	Pyruvat Carboxylase	ATP + Pyruvat + HCO_3^- = ADP + Phosphat + Oxaloacetat
6.4.1.2	Acetyl-CoA Carboxylase	ATP + Acetyl-CoA + HCO_3^- = ADP + Phosphat + Malonyl-CoA
6.4.1.3	Propionyl-CoA Carboxylase	ATP + Propanoyl-CoA + HCO_3^- = ADP + Phosphat + (S)-Methylmalonyl-CoA
6.4.1.4	Methylcrotonoyl-CoA Carboxylase	ATP + 3-Methylcrotonoyl-CoA + HCO_3^- = ADP + Phosphat + 3-Methylglutaconyl-CoA

Während die Reaktionen der jeweils gleichen Hauptklasse deutliche Parallelen zueinander aufweisen, bestehen zwischen den Hauptklassen keine offensichtlichen Ähnlichkeiten. Auffallend ist allein, daß fünf der sieben dargestellten Reaktionen unter Spaltung von ATP verlaufen, wobei ADP entsteht. Die Reaktion der UDP-Sulfoquinovose Synthase verläuft indes unter Beteiligung von

UDP-Glucose, deren Substruktur ADP ähnelt. Dies könnte ein mögliches Indiz dafür sein, daß eine gemeinsame Nucleosiddiphosphat-Bindungsstelle Ursache für die Gruppierung der Enzyme ist. Dies könnte wiederum auch die große Zahl der Oxidoreduktasen im Cluster erklären, da diese oftmals NADH bzw. NADPH als Cofaktor benötigen, deren Substrukturen ebenfalls ein Nucleosiddiphosphat beinhalten. Wie läßt sich vor diesem Hintergrund jedoch die Gruppierung der ATP-unabhängigen DNA-Topoisomerase erklären?

Die Wirkungsweise der beiden DNA-Topoisomerasen ist sehr ähnlich. Sie verändern die Topologie von DNA-Molekülen, indem sie superhelikale DNA relaxieren bzw. entspannte DNA in superhelikale DNA überführen. Sie bewirken dies über eine vorübergehende Öffnung des Phosphodiester-Bandes, gefolgt von der Passage eines DNA-Stranges durch die Lücke und nachfolgender Wiederherstellung der Phosphodiester-Bindung. Der Unterschied zwischen beiden Enzymen besteht darin, daß die Reaktion der ATP-unabhängige DNA-Topoisomerase (Topoisomerase I) unter vorübergehender Spaltung eines der beiden DNA-Stränge verläuft, während die ATP-abhängige DNA-Topoisomerase (Topoisomerase II) zwischenzeitlich beide Stränge spaltet. Die freie Energie der ATP-Spaltung wird in diesem Fall für die Konformationsänderung benötigt, die das Enzym während des Reaktionszyklus durchläuft [79].

Die ähnliche Wirkungsweise spiegelt sich auch in der Ähnlichkeit ihrer Sequenzen wider. So können Sequenzen beider Enzyme erstmals bei einem Grenzwert von 10^{-50} in einem gemeinsamen Cluster nachgewiesen werden. Es kann daher von einem gemeinsamen evolutionären Ursprung beider Enzyme ausgegangen werden. Es besteht somit die Möglichkeit, daß auch die ATP-unabhängigen DNA-Topoisomerase über eine entsprechende Nucleosiddiphosphat-Bindungsstelle verfügt, diese für die Enzymfunktion aber nicht mehr benötigt wird. Gestützt wird diese Vermutung durch die Beobachtung, daß die ATP-abhängige Topoisomerase II unter bestimmten Vorraussetzungen auch als ATP-unabhängige Topoisomerase vom Typ I fungieren kann [79].

Die exemplarische Betrachtung der sieben ausgewählten Enzymklassen eröffnet somit die Möglichkeit, daß auch dieses bei 10^{-2} erhaltene, sehr heterogene Cluster noch auf einer entwicklungsgeschichtlichen Verwandtschaft der gruppierten Sequenzen beruht. Ohne umfassende Analyse aller am Cluster beteiligten Enzyme bleibt dieses jedoch Spekulation, da unklar ist, inwieweit sich die Eigenschaften auch auf die übrigen Mitglieder des Clusters übertragen lassen.

Im Verlauf der Auswertung ergaben sich jedoch noch weitere Hinweise darauf, daß auch die bei 10^{-2} generierten Cluster auf einer evolutionären wie auch mechanistischen Verwandtschaft der gruppierten Enzyme beruhen.

3.6.3 Exemplarische Betrachtung ausgewählter Cluster

Die bei der Clusteranalyse erhaltenen Gruppierungen geben Auskunft darüber, welche Enzymsequenzen einen gemeinsamen phylogenetischen Ursprung besitzen und auf welchen Bereich ihrer Sequenzen sich ihre Verwandtschaft bezieht. Die Ähnlichkeit ihrer Reaktionsmechanismen muß hingegen noch eruiert werden.

Da sich die Analyse der erhaltenen Sequenzcluster aufgrund der dazu notwendigen Recherchen nicht automatisieren läßt, war es nicht möglich, alle in Frage kommenden Cluster auszuwerten. Es wurden daher einige Cluster für eine manuelle Auswertung ausgewählt. Die Auswahl erfolgte aufgrund der Größe und Zusammensetzung der generierten Cluster und konzentriert sich auf solche Enzympaarungen, deren EC-Klassifikation sich bereits in der Hauptklasse voneinander unterscheiden. Überdies war es für die Untersuchung hilfreich, wenn Strukturdaten zu den jeweiligen Enzymklassen zur Verfügung standen. Im folgenden werden einige der betrachteten Cluster exemplarisch dargestellt.

3.6.3.1 Klasse II Aldolasen

In der nachfolgenden Tabelle 3-13 ist die Sequenzzusammensetzung eines der betrachteten Cluster dargestellt. Zu jeder Enzymklasse ist die Zahl der im Cluster sowie der im gesamten Datensatz enthaltenen Sequenzen angegeben.

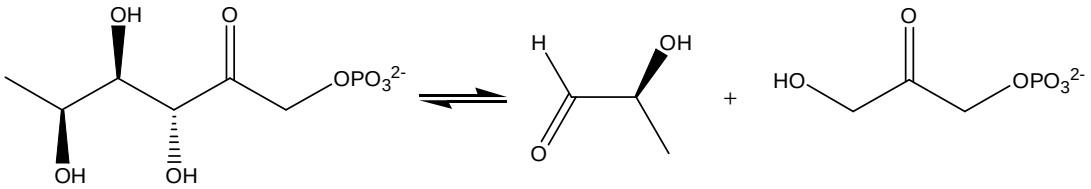
Tabelle 3-13: Übersicht über die Sequenzzusammensetzung des Clusters

EC-Klasse	Enzym	Sequenzzahl			
		Cluster			Datensatz
		10^{-10}	10^{-5}	10^{-2}	
4.1.2.17	L-Fuculose-phosphat-Aldolase	15	15	15	15
4.1.2.19	Rhamnulose-1-phosphat-Aldolase	0	1	6	6
5.1.3.4	L-Ribulose-phosphat 4-Epimerase	12	12	12	12

Das Cluster enthält je nach Grenzwert zwischen 27 und 33 Sequenzen, die zwei bzw. drei unterschiedlichen Enzymklassen angehören. Zwei von ihnen klassifizieren Aldolasen, die zur Hauptklasse der Lyasen zählen. Die dritte beschreibt eine Epimerase, die zur Hauptklasse der Isomerasen gehört. Die unterschiedliche Sequenzzusammensetzung bei den verschiedenen Grenzwerten zeigt, daß die Sequenzen der L-Fuculose-phosphat-Aldolasen und der L-Ribulose-phosphat 4-Epimerasen einander ähnlicher sind als die der beiden Aldolasen.

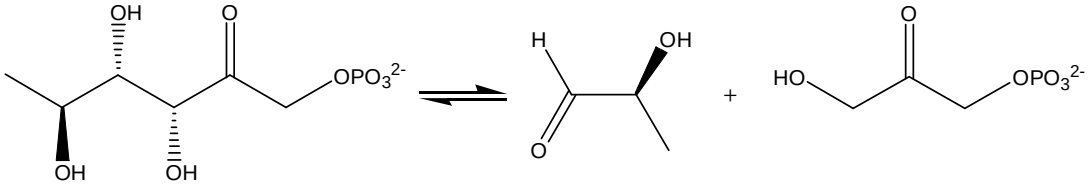
Trotz ihrer nur geringen Sequenzähnlichkeit katalysieren die beiden Aldolasen nahezu identische Reaktionen. Die L-Fuculose-phosphat-Aldolase (FucA) ist ein Enzym aus dem Fucose-Stoffwechsel und katalysiert die reversible Spaltung von L-Fuculose-1-phosphat in L-Lactaldehyd und Dihydroxyacetonphosphat (DHAP) [80, 81] (vgl. Abbildung 3-11).

Die Rhamnulose-1-phosphat-Aldolase (RhuA) ist am Katabolismus der L-Rhamnose beteiligt und katalysiert die reversible Spaltung von L-Rhamnulose-1-phosphat in L-Lactaldehyd und Dihydroxyacetonphosphat [82] (vgl. Abbildung 3-12). Bei ihren Substraten handelt es sich somit um Epimere, die sich lediglich in der Konfiguration ihres C4-Atoms voneinander unterscheiden.



L-Fuculose-1-phosphat (S)-Lactaldehyd Dihydroxyacetonphosphat

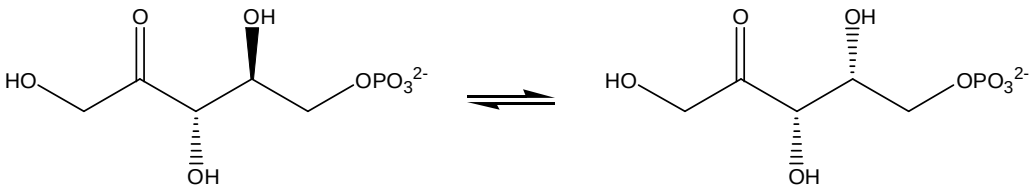
Abbildung 3-11: Schematische Darstellung der von L-Fuculose-phosphat-Aldolase katalysierten Reaktion



L-Rhamnulose-1-phosphat (S)-Lactaldehyd Dihydroxyacetonphosphat

Abbildung 3-12: Schematische Darstellung der von Rhamnulose-1-phosphat-Aldolase katalysierten Reaktion

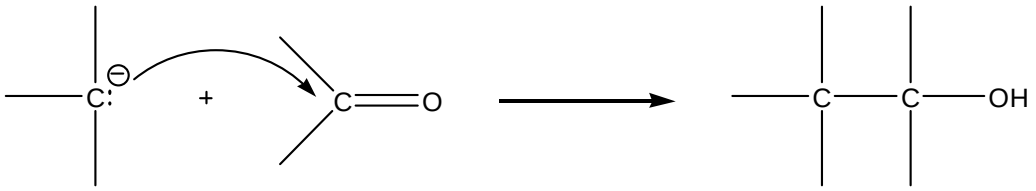
Das Enzym L-Ribulose-phosphat 4-Epimerase (RibE) ist am dritten Schritt des L-Arabinose-Katabolismus beteiligt und katalysiert die gegenseitige Umwandlung von L-Ribulose 5-phosphat und D-Xylulose 5-phosphat [83].



L-Ribulose 5-phosphat D-Xylulose 5-phosphat

Abbildung 3-13: Schematische Darstellung der von L-Ribulose-phosphat 4-Epimerase katalysierten Reaktion

Aldolase katalysierte Reaktionen führen zur Spaltung oder zur Bildung von C-C-Bindungen und bilden die Grundlage sowohl des katabolen als auch des anabolen Stoffwechsels. Vom Standpunkt der Synthese aus betrachtet, verlaufen solche Reaktionen unter Angriff eines nucleophilen Carbanions auf ein elektrophiles C-Atom, im vorliegenden Fall das Carbonyl-C-Atom von Aldehyden.



Beim Angriff an diesen elektrophilen Zentren müssen stabilisierte Carbanionen erzeugt werden. Als Zwischenprodukt entsteht zunächst das durch bessere Elektronendelokalisierung stabilisierte Enolat (vgl. Abbildung 3-14a). Dessen Stabilisierung wird durch das Enzym noch verstärkt. Entsprechend des dabei verwendeten Mechanismus werden zwei Klassen von Aldolasen unterschieden [84]. Bei Klasse I Aldolasen, die bei Tieren und Pflanzen vorkommen, verläuft die Reaktion unter Bildung eines Iminium-Kations, d.h. einer protonierten Schiff-Base (kovalente Katalyse; vgl. Abbildung 3-14b). Bei Klasse II Aldolasen, die bei Pilzen, Algen und einigen Bakterien vorkommen, polarisiert ein zweiwertiges Kation, üblicherweise Zn^{2+}, den Carbonylsauerstoff des Substrats, wodurch das Enolat-Zwischenprodukt der Reaktion stabilisiert wird (Metall-Ionen-Katalyse; vgl. Abbildung 3-14c). Die beiden im Cluster enthaltenen Enzyme gehören zu den Klasse II Aldolasen.

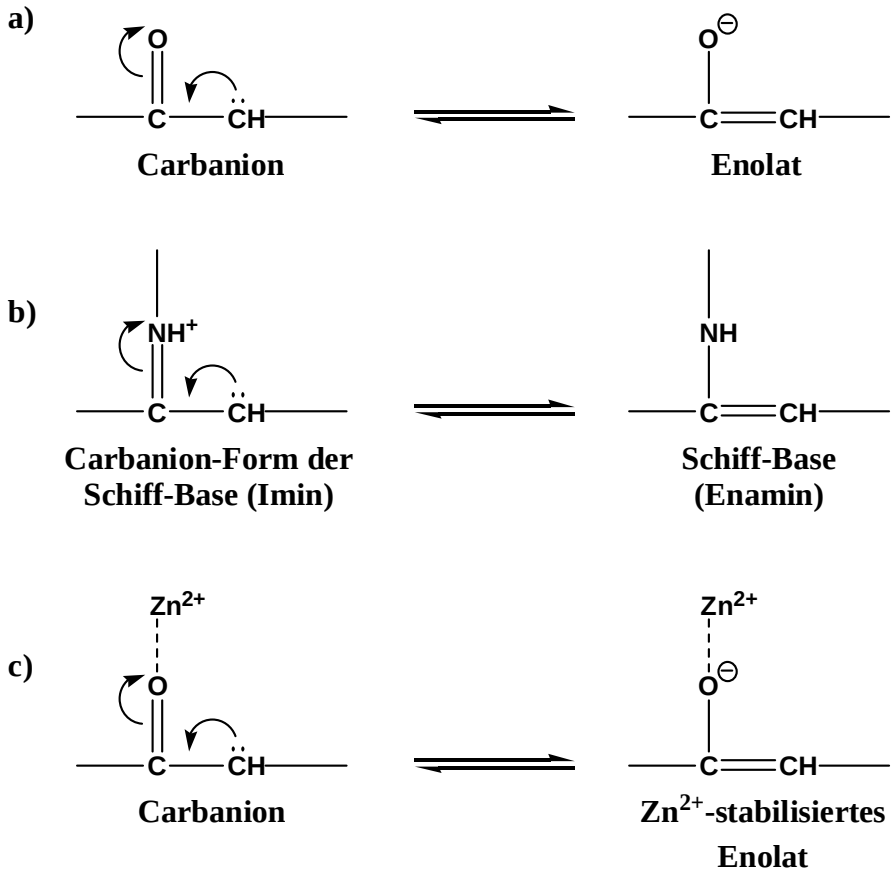


Abbildung 3-14: Stabilisierung von Carbanionen: (a) Carbanionen in Nachbarschaft zu einer Carbonylgruppe werden durch die Bildung von Enolaten stabilisiert. (b) Carbanionen in Nachbarschaft zu protonierten Iminen (Schiff-Basen) werden durch die Bildung von Enaminen stabilisiert. (c) Metall-Ionen stabilisieren Carbanionen neben Carbonylgruppen durch elektrostatische Stabilisierung des Enolats.

Die Klasse II Aldolasen müssen je nach Faltung und Mechanismus weiter unterteilt werden [85]. Das vorherrschende Strukturmotiv der Aldolasen ist das $(\beta\alpha)_8$-Faß. Dieses auch unter dem Namen *TIM-barrel* bekannte Motiv besteht aus einem achtsträngigen parallelen β-Faltblatt in Form eines Fasses, das auf der Außenseite von acht α-Helices umgeben ist. Die Faltungsstruktur der beiden im Cluster befindlichen Aldolasen unterscheidet sich jedoch von diesem. Sie weisen ein dreilagiges $\alpha/\beta/\alpha$-Strukturmotiv auf, das aus einem beidseitig von α-Helices umgebenden neunsträngigen vorwiegend antiparallelen β-Faltblatt besteht [86]. Die beiden Aldolasen besitzen somit das gleiche Strukturmotiv wie die L-Ribulose-phosphat 4-Epimerasen [87].

Um die Ähnlichkeit zwischen den Strukturen der verschiedenen EC-Klassen zu quantifizieren, wurden mittels des Programms Protein3Dfit [88] Strukturalignments erstellt. Standen mehrere Strukturen zu einer Enzymklasse zur Verfügung, so wurde jeweils die mit der besten Auflösung zur Betrachtung ausgewählt. Zwei Faltungen werden als strukturell ähnlich angesehen, wenn sich mindestens 40% der korrespondierenden C_α-Atome innerhalb eines Abstandes von maximal 1,8 Å zueinander befinden. Tabelle 3-14 enthält eine Übersicht der Ergebnisse.

Tabelle 3-14: Übersicht über die strukturelle Ähnlichkeit zwischen den Enzymen der verschiedenen EC-Klassen.

EC-Klasse	Swiss-Prot Nummer	PDB-Struktur	1E4B	1OJR	1K0W
4.1.2.17	**P11550**	**1E4B**	-		
4.1.2.19	**P32169**	**1OJR**	57,28%	-	
5.1.3.4	**P08203**	**1K0W**	74,76%	46,64%	-

Die Strukturalignments belegen die evolutionäre Verwandtschaft der miteinander gruppierten Sequenzen und bestätigen so die mittels Sequenzähnlichkeit getroffene Einteilung der Enzyme. Wie der Tabelle zu entnehmen ist, bestehen die größten strukturellen Übereinstimmungen zwischen FucA und RibE. RhuA zeigt dagegen zu RibE nur eine relativ geringe Ähnlichkeit. Somit ergibt sich anhand der Strukturvergleiche das gleiche Ergebnis wie mittels der Clusterung.

Die Röntgenstruktur der FucA aus *E. coli* wurde von Dreyer & Schulz in zwei Kristallformen sowie im Komplex mit Phosphoglykolohydroxamat aufgeklärt [87, 89, 90]. Phosphoglykolohydroxamat ähnelt dem Endiolat-Übergangszustand von Dihydroxyacetonphosphat nach Deprotonierung am Kohlenstoffatom C3 und wirkt als kompetitiver Inhibitor [85]. Daraus wurden Einzelheiten über die Struktur des katalytischen Zentrums und den Reaktionsmechanismus abgeleitet.

Das Enzym ist ein Homotetramer. Die vier aktiven Zentren des FucA-Tetramers werden durch die katalytisch aktiven Zink-Ionen markiert und liegen in einer Spalte in der Kontaktregion (*interface*) zwischen zwei benachbarten Untereinheiten. In der unligierten Struktur wird das katalytische Zink-Ion annähernd tetraedrisch von drei Histidin-Seitenketten (His92, His94 und His155)

und der Carboxylatgruppe von Glu73 koordiniert. Darüber hinaus ragt Tyr113 von der benachbarten Untereinheit ins aktive Zentrum und nähert sich mit seiner phenolischen Hydroxylgruppe dem Zink-Ion bis auf einen Abstand von 3.5 Å an. In der Inhibitorstruktur wird Glu73 durch die Bindung des Dihydroxyacetonphosphat-Analogons Phosphoglykolohydroxamat aus der Koordinationssphäre verdrängt und das Zink-Ion annähernd quadratisch pyramidal koordiniert. Die Phosphatgruppe des Phosphoglykolohydroxamat wird in einer aus den Resten Gly28, Asn29, Thr43, Gly44, Ser71 und Ser72 gebildeten Tasche durch Wasserstoffbrücken gebunden.

Aus diesen strukturellen Daten wurde von Dreyer & Schulz (1996) ein Reaktionsmechanismus abgeleitet, wonach je nach Richtung der katalysierten Reaktion Glu73 und Tyr113 als Säure bzw. Base fungieren [89]. Aufgrund von kinetischen Untersuchungen sowie Röntgenstrukturen zahlreicher FucA-Mutanten wurde der Vorschlag von Joerger et al. (2000) dahingehend abgewandelt, daß Glu73 die einzige Säure und Base der katalysierten Reaktion ist [85] (vgl. Abbildung 3-15). Die Funktion von Tyr113 liegt vermutlich in der Positionierung der Aldehyd-Komponente und somit in der Kontrolle der Enantio- und Diastereoselektivität der Reaktion.

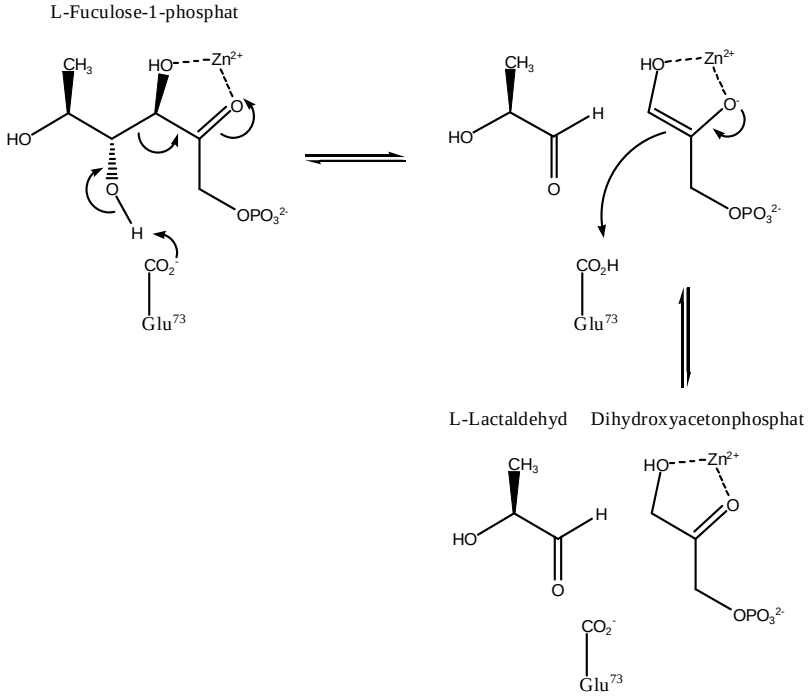


Abbildung 3-15: Reaktionsmechanismus der L-Fuculose-1-phosphat Aldolase [89]. Die Positionsangabe der Aminosäure bezieht sich auf die Sequenz aus *E. coli*.

Der Reaktion der Rhamnulose-1-phosphat-Aldolase verläuft auf ähnliche Weise, mit Ausnahme einer Drehung der Aldehydgruppe um 180°, welche die beiden Epimere Rhamnulose und Fuculose voneinander unterscheidet [86].

Enzymatische Isometrisierungen spielen eine wichtige Rolle im Kohlenhydrat-Stoffwechsel von Organismen. Eine Untergruppe der Isomerasen sind die Epimerasen, die eine Umkehr der Stereokonfiguration an einem asymmetrischen Kohlenstoffatom katalysieren.

Die L-Ribulose-5-phosphat-4-Epimerase aus *E. coli* ist eines von drei Enzymen des Arabinose-Operons [91], die die Nutzung von Arabinose als Energiequelle ermöglichen. RibE katalysiert die reversible Umwandlung der offenkettigen Pentosen L-Ribulose-5-phosphat und D-Xylulose-5-phosphat. Die Epimerisierung erfolgt dabei am Kohlenstoffatom C4, das aufgrund des Fehlens einer direkt benachbarten Carbonylfunktion nur eine geringe Azidität aufweist. Eine Epimerisierung durch einfache Deprotonierung/Reprotonierung unter Ausbildung eines cis-2,3-Endiolat-Zwischenproduktes, wie es für die D-Ribulose-5-phosphat-3-Epimerase aus den Chloroplasten von Kartoffel postuliert wurde [92], kann somit ausgeschlossen werden. Da für die Katalyse NAD^+ als Cofaktor nicht benötigt wird [93], kann ein Mechanismus unter Hydridübertragung durch aufeinander folgende Oxidation und Reduktion des Substrats analog zur NAD^+-abhängigen UDP-Galaktose-4-Epimerase [94] ebenfalls ausgeschlossen werden.

Aufgrund der Abhängigkeit der katalysierten Reaktion von zweiwertigen Metallionen wurden von Deupree & Wood (1972) zwei weitere mögliche Reaktionsmechanismen für die RibE vorgeschlagen [95]. Die Epimerisierung der Substrate kann demnach entweder durch eine Dehydratisierung unter Ausbildung einer Doppelbindung zwischen den Kohlenstoffatomen C3 und C4 mit anschließender Rehydratisierung oder aber durch einen Retroaldol-/Aldol-Mechanismus unter intermediärer Spaltung der C3-C4-Kohlenstoffbindung erfolgen. Im Fall einer Dehydratisierung müssen allerdings das zunächst am Kohlenstoffatom C3 abstrahierte Proton sowie das anschließend eliminierte Hydroxidion erneut an die Doppelbindung und zwar unter Umkehrung der Stereochemie am Kohlenstoffatom C4 addiert werden, da das ^3H markierte D-Xylulose-5-phosphat die Markierung im Verlauf der Epimerisierung nicht

verliert und ein Einbau von Isotopen aus ^{3}H- oder ^{18}O-markiertem Solvens nicht stattfindet [96].

Die Verifizierung des Retroaldol-/Aldol-Mechanismus gelang Johnson & Tanner (1998) durch experimentellen Nachweis der beiden Teilreaktionen [97]. So konnte gezeigt werden, daß RibE ausgehend von den postulierten Zwischenprodukten Dihydroxyaceton und Glycolaldehydphosphat in geringem Umfang die Kondensation zu L-Ribulose-5-phosphat und D-Xylulose-5-phosphat katalysiert. Die Konservierung von Aminosäuren sowohl in der Koordinationssphäre des Zink-Ions als auch im Bereich der Phosphatbindungstasche [98] sowie kinetische Untersuchungen von Mutanten konservierter Reste deuten zusätzlich auf eine mechanistische Verwandtschaft zu den obengenannten Klasse II Aldolasen hin [97, 99, 100].

Jede RibE-Untereinheit enthält ein Zink-Ion, das von drei Histidin- Seitenketten (His95, His97 und His171) koordiniert wird und das aktive Zentrum des Enzyms im *interface* zwischen zwei benachbarten Untereinheiten markiert. Die für die Katalyse entscheidenden Reste wurden durch ortsgerichtete Mutagenese und anschließende kinetische Untersuchungen aufgeklärt [99, 100]. Aus den kinetischen Daten der RibE-Mutante D76N wurde geschlossen, daß Asp76 nicht wie zuvor von Johnson & Tanner (1998) vermutet dieselbe Funktion wie Glu73 in der FucA als einzige Säure und Base der katalysierten Reaktion zukommt. Vielmehr wurde der mit Abstand größte Einfluß auf die katalytische Aktivität bei den Mutanten D120N und Y229F beobachtet, die im Vergleich zum Wildtyp eine 3000-fach bzw. 1000-fach geringere Aktivität aufwiesen. Während Asp120 und das im Bereich des beweglichen C-terminalen Kettenendes liegende Tyr229 innerhalb von RibE-Sequenzen strikt konserviert sind, zeigen die Sequenzen der Klasse-II-Aldolasen an den entsprechenden Positionen unterschiedliche Aminosäurereste [98]. Weiterhin konnte anhand der äußerst geringen Substratsaffinität der Mutanten N28A und K42M bewiesen werden, daß die Phosphatgruppe der RibE-Substrate in der zur FucA homologen Phosphatbindungstasche gebunden wird. Im Vergleich zum FucA-Substrat L-Fuculose-1-phosphat, das die Phosphatgruppe am Kohlenstoffatom C1 trägt, werden die RibE-Substrate mit ihrer Phosphatgruppe am C5-Atom folglich in Bezug auf die Kohlenstoffatome C2 und C3 in umgekehrter Orientierung an das Zink-Ion koordiniert [101].

Unter Berücksichtigung der biochemischen Daten wurden die natürlichen Substrate der RibE in das aktive Zentrum modelliert und ein Reaktionsmechanismus vorgeschlagen [100, 102] (vgl. Abbildung 3-16). Demnach werden L-Ribulose-5-phosphat und D-Xylulose-5-phosphat über die Phosphatgruppe in der Phosphatbindungstasche sowie über den Carbonylsauerstoff am C2-Atom und den Hydroxylsauerstoff am C3-Atom an das Zink-Ion gebunden.

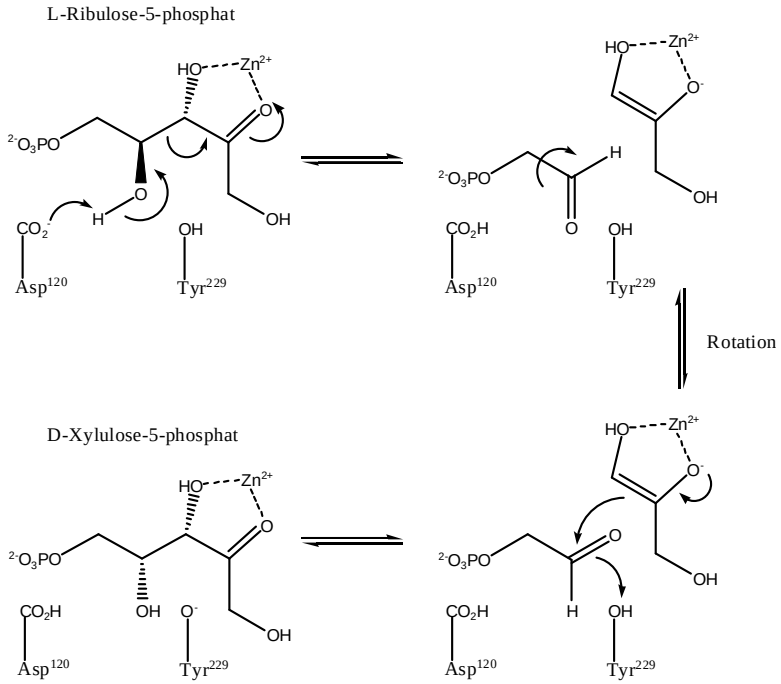


Abbildung 3-16: Reaktionsmechanismus der L-Ribulose-phosphat 4-Epimerase [100, 102]. Die Positionsangaben der Aminosäuren beziehen sich auf die Sequenz aus *E. coli*.

Als katalytisch aktive Säure bzw. Base fungieren Asp120 und Tyr229, die beide aus der benachbarten Untereinheit stammen. Ausgehend von gebundenem D-Xylulose-5-phosphat wird dieses von Asp120 an der Hydroxylgruppe des C4-Atoms deprotoniert und die Bindung zwischen den Kohlenstoffatomen C3 und C4 gebrochen. Die Stabilisierung des dabei entstehenden Endiolats erfolgt wie bei den Klasse-II-Aldolasen durch Komplexierung an das Zink-Ion. Im Gegensatz zu den Klasse-II-Aldolasen werden die Spaltprodukte allerdings nicht freigesetzt, sondern bleiben durch das Enzym vom Solvens abgeschirmt. Nach Rotation der Aldehydcarbonylgruppe erfolgt ein nukleophiler Angriff des

Endiolats unter erneuter Ausbildung der C3-C4-Kohlenstoffbindung. Die Reprotonierung des so gebildeten L-Ribulose-5-phosphat erfolgt nach Fixierung der C-terminalen Reste durch Tyr229, das im Vergleich zu Asp120 auf der gegenüber liegenden Seite des Substrats positioniert ist.

Zusammenfassen ist festzuhalten, daß die Enzyme RhuA, FucA und RibE die Spaltung einer C-C-Bindung katalysieren, wobei im Fall der RibE allerdings die Produkte dieser Reaktion nicht freigesetzt werden, sondern im Anschluß an die Rotation der gebildeten Aldehydgruppe um 180° die erneute Ausbildung der zuvor gespaltenen C-C-Bindung stattfindet. Den drei Enzymen ist weiterhin gemeinsam, daß ein katalytisch aktives Zinkion zur Stabilisierung des durch die Spaltung der C-C-Bindung gebildeten Endiolations verwendet wird. Auch in Bezug auf die Aktivierung der nicht-aziden 4-Hydroxylgruppe in β-Stellung zur Carbonylgruppe in den als Substraten dienenden Phosphoketosen bestehen Übereinstimmungen, da für diese Aufgabe eine katalytisch aktive Base von den Enzymen zur Verfügung gestellt werden muß. Im Falle der Aldolasen müssen die Enzyme für die Protonierung des gebildeten Endiolations bzw. im Fall der Epimerase für die Reprotonierung zur Ausbildung der 4-epimeren Hydroxylgruppe zusätzlich auch eine katalytisch aktive Säure zur Verfügung stellen. Grundsätzlich ist es dabei möglich, daß diese katalytisch aktiven Säuren und Basen entweder aus ein und derselben Aminosäure oder aber aus zwei verschiedenen Aminosäuren bestehen. Die erste dieser beiden Möglichkeiten wird in der FucA mit dem Glu73 gefunden [85], die andere Möglichkeit ist vermutlich in der RibE durch Asp120 und Tyr229 realisiert [100].

3.6.3.2 Klasse II Glutamin Amidotransferasen

Tabelle 3-15 enthält die Sequenzzusammensetzung eines weiteren zur Betrachtung ausgewählten Clusters. Erneut sind zu jeder Enzymklasse die Gesamtzahl der im Datensatz und die Zahl der bei drei verschiedenen Grenzwerten im Cluster enthaltenen Sequenzen angegeben.

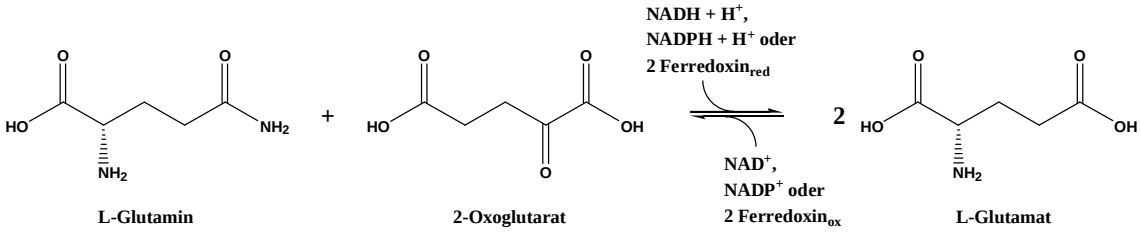
Tabelle 3-15: Sequenzzusammensetzung des ausgewählten Clusters

EC-Klasse	Enzym	Sequenzzahl			
		Cluster			Datensatz
		10^{-10}	10^{-5}	10^{-2}	
1.4.1.13	Glutamat Synthase (NADPH)	0	1	13	50
1.4.1.14	Glutamat Synthase (NADH)	0	0	1	2
1.4.7.1	Glutamat Synthase (Ferredoxin)	0	0	4	17
2.4.2.14	Amidophosphoribosyltransferase	65	66	66	66
2.6.1.15	Glutamin-pyruvat Transaminase	0	1	1	2
2.6.1.16	Glutamin-fructose-6-phosphat Transaminase (isomerisierend)	117	117	117	120
6.3.5.4	Asparagin Synthase (Glutamin-hydrolysierend)	1	52	52	54

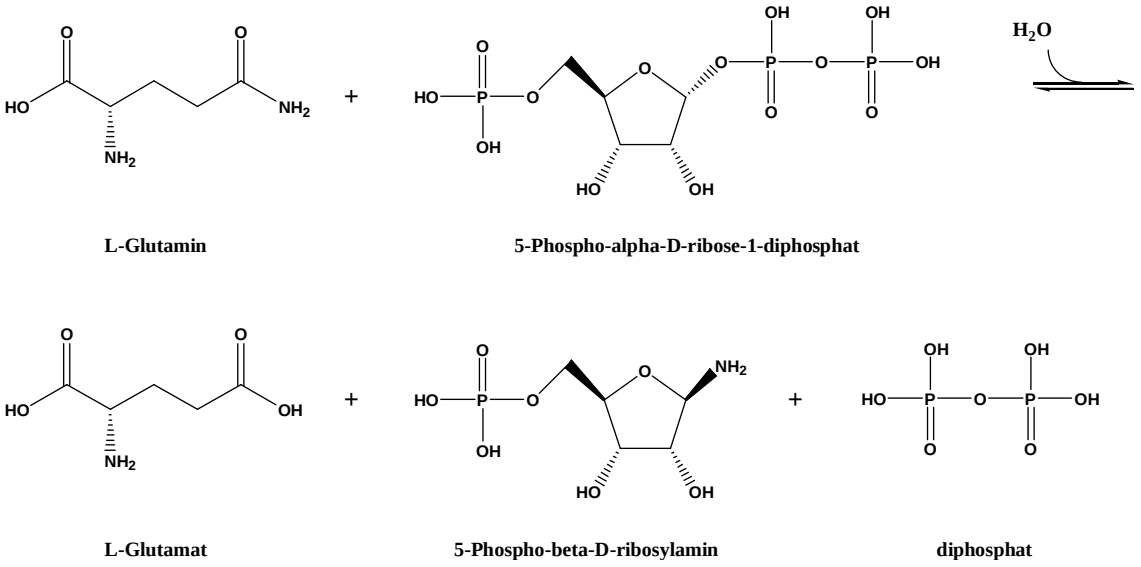
Von allen betrachteten Clustern zeigt dieses bei den verschiedenen Grenzwerten die größte Änderung in der Zusammensetzung der Sequenzen. Bei einem Grenzwert von 10^{-10} enthält das Cluster 183 Sequenzen aus drei verschiedenen Enzymklassen. Diese entstammen den Hauptklassen der Transferasen und Ligasen, wobei die Ligasen lediglich mit einer Sequenz im Cluster vertreten sind. Beim Grenzwert von 10^{-5} kommen Sequenzen aus zwei weiteren Enzymklassen hinzu, von der eine der Hauptklasse der Oxidoreduktasen angehört. Beim höchsten Grenzwert enthält das Cluster schließlich 255 Sequenzen aus insgesamt sieben verschiedenen Enzymklassen, die drei verschiedenen Hauptklassen entstammen.

Die Glutamat Synthase ist eines der Schlüsselenzyme bei der Aufnahme von Ammoniak in Bakterien, Algen und Pflanzen. Es gibt drei Formen des Enzyms, die Ferredoxin, NADH bzw. NADPH als Elektronendonor verwenden [103]. Alle drei Formen katalysieren die reduktive Transaminierung der Aminogrup-

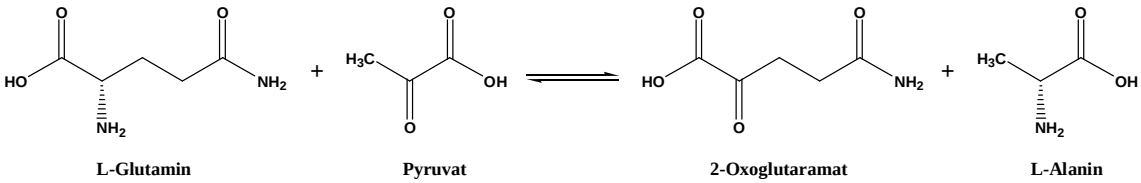
pe von L-Glutamin zu 2-Oxoglutarat, bei der zwei Moleküle L-Glutamat entstehen [104].



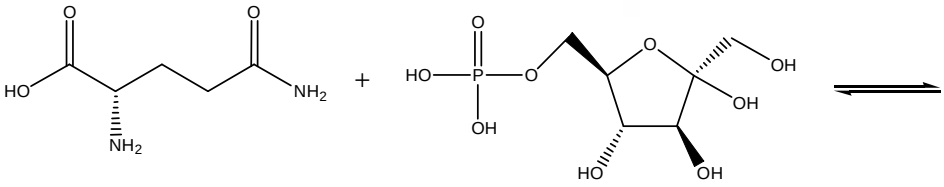
Die Amidophosphoribosyltransferase katalysiert den ersten Schritt in der Purin-Biosynthese: die Übertragung der Aminogruppe von L-Glutamin auf 5-Phospho-α-D-ribose-1-diphosphat unter Bildung von 5-Phospho-β-D-ribosylamin [105].



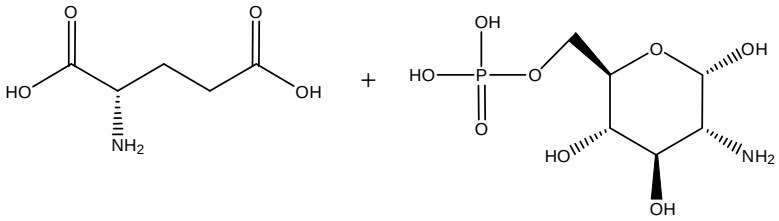
Das Enzym Glutamin-pyruvat Transaminase katalysiert die Pyridoxal-5′-phosphat-abhängige Transaminierung der Aminogruppe von L-Glutamin zu Pyruvat, bei der L-Alanin und 2-Oxoglutaramat entstehen [106].



Die Glutamin-fructose-6-phosphat Transaminase katalysiert eine Schlüsselreaktion in der Hexosamin-Synthese: die Bildung von D-Glycosamin-6-phosphat aus D-Fructose-6-phosphat und L-Glutamin [107].

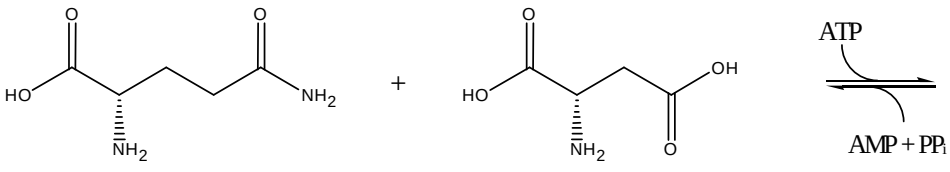


L-Glutamin D-Fructose-6-phosphat

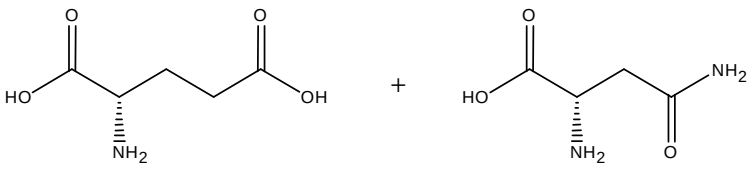


L-Glutamat D-Glucosamin-6-phosphat

Die Synthese von Asparagin wird durch die Asparagin Synthase katalysiert, die ATP als Cofaktor benötigt. Die Reaktion erfolgt über die Bildung von β-Aspartyl-adenosylmonophosphat, das anschließend eine nucleophile Acylsubstitution mit einer aus L-Glutamin stammenden Aminogruppe eingeht [108].



L-Glutamin L-Aspartat



L-Glutamat L-Asparagin

Alle im Cluster zusammengefaßten Enzyme verwenden Glutamin als eines ihrer Substrate. Glutamin ist bei vielen Biosynthesen der Aminogruppen-Donor. Die Aminogruppe von Glutamin ist die Hauptstickstoffquelle für die Synthese verschiedener Biomoleküle wie zum Beispiel Aminosäuren, Purin- und Pyrimidin-Nukleotide und Coenzyme [109]. Bislang sind 16 GATasen (GAT: *glutamine amide transfer*) bekannt, deren funktionelle Gemeinsamkeit darin besteht, daß Glutamin hydrolysiert und die Aminogruppe in Form von nascentem Ammoniak auf ein Akzeptorsubstrat übertragen wird [110].

GATasen sind modulare Enzyme [111], die aus einer Glutaminase-Domäne (GAT-Domäne) und einer Synthase-Domäne aufgebaut sind. Die Synthase-Domänen verschiedener GATasen unterscheiden sich in ihren Substraten, der Reaktion, die sie katalysieren, sowie ihrer Faltungstopologie [110]. Die Glutaminase-Domänen wurden aufgrund der katalytisch essentiellen Aminosäuren und konservierter Aminosäuremotive in zwei Klassen eingeteilt und diese Einteilung auf die GATase-Holoenzyme übertragen [109, 112]. Aufgrund der fehlenden Sequenzhomologie zwischen Klasse-I- und Klasse-II-GATasen sowie unterschiedlicher Proteinfaltungstopologien kann davon ausgegangen werden, daß die beiden GATase-Klassen nicht miteinander verwandt sind [109].

Die Klasse I der GATasen besitzen drei für die Katalyse essentiellen Aminosäuren: Cystein, Histidin und Glutamat. Diese Aminosäuren bilden eine katalytische Triade wie sie bei verschiedenen hydrolytischen Enzymen vorkommen [113]. Die fehlende Sequenzhomologie zu anderen Proteinfamilien zeigt, daß die Klasse-I-GATasen nicht mit anderen bekannten Proteinfamilien verwandt sind [109].

Die Klasse-II-GATasen sind Mitglieder der großen Familie der „Ntn"-Hydrolasen [114]. Zu dieser Familie gehören zahlreiche hydrolytische Enzyme, wie z. B. das Proteasom und die Penicillin-Acylase. „Ntn" steht für N-terminales Nucleophil und deutet damit auf eine Gemeinsamkeit dieser Enzyme. Das N-terminale Nucleophil ist im Proteasom ein Threonin, in der Penicillin-Acylase ein Serin und in den Klasse-II-GATasen ein Cystein [115]. Der erste Schritt in der Substrathydrolyse besteht stets in dem Angriff dieses Nucleophils.

Die im Cluster zusammengefaßten Enzyme gehören zu den Klasse-II-GATasen. Dies spiegelt sich in einem gemeinsamen Sequenzmotiv (PROSITE-Motiv GATASE_TYPE_II (PS00443)) wider, das allen Sequenzen des Clusters gemeinsam ist und welches das katalytisch aktive Cystein beinhaltet. Abbildung 3-17 zeigt ein multiples Sequenzalignments der entsprechenden Region mit je einem Vertreter aus jeder Enzymklassen.

```
Q8PY97|1.4.1.13   12  RVIKM-----------------------------------CGIIGFID---RTKSRMDG   32
P19252|6.3.5.4     1  ---------------------------------------CGILAVLG--CSDPSRAKR   17
Q8DEF3|2.6.1.16    1  ---------------------------------------CGIVGAVA------QRDVA   13
Q8FMM7|2.4.2.14   16  VNYPYDD-----------------------HNEQSPQEECGVFGVWA-----PGEEVA   45
O08339|2.6.1.15   23  GLYKAED-----------------------------EHASCGV-GLVVSISGTPSRKVV   51
Q03460|1.4.1.14   61  QLWESGGLGRLPKLRVAVKSSFSAVPDKPMGLYDPAFDKDSCGV-GFVAELNGQSSRKTV  119
Q7U4D7|1.4.1.7     8  TVWPYSD--------------------SAAPEAVAGEKDACGV-GFLAQLSGETSHWVL   45
                                                    **: .
```

Abbildung 3-17: Ausschnitt des multiplen Sequenzalignments zum betrachteten Cluster. Die Region des Alignments, die das von PROSITE definierte Sequenzmotiv (P00443) mit dem katalytisch aktiven Cystein enthält, ist hervorgehoben.

Obwohl die GAT-Domänen der Klasse-II-GATasen nur eine katalytisch essentielle Aminosäure besitzen, die GAT-Domänen der Klasse-I-GATasen dagegen drei, verläuft die Glutaminhydrolyse in allen GATasen nach einem ähnlichen Schema ab [110, 117, 118]:

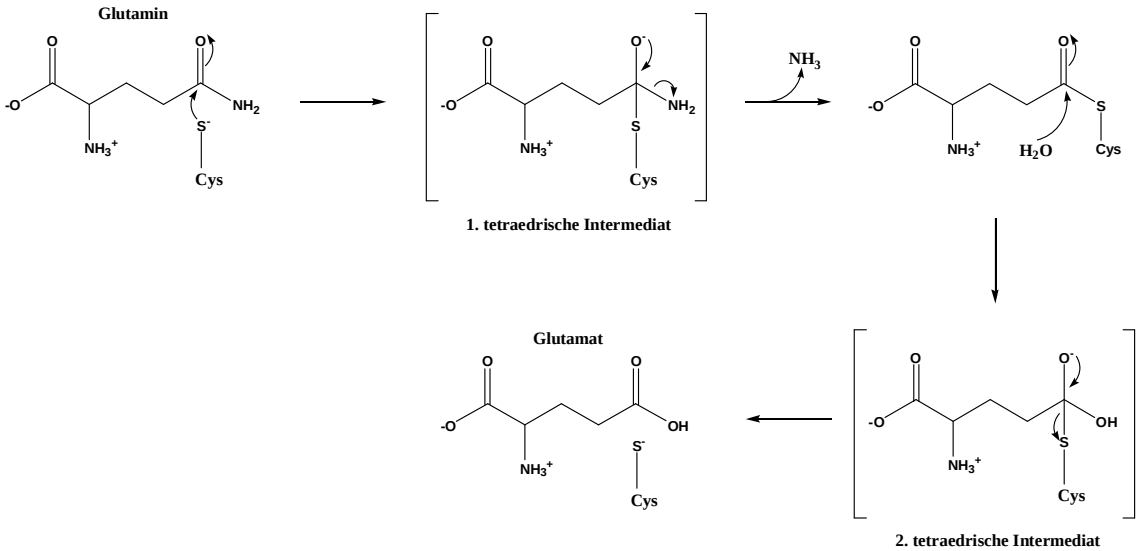


Abbildung 3-18: Reaktionsmechanismus der Glutaminase-Reaktion in Klasse-II-GATasen (nach van den Heuvel *et al.* [118])

Der nukleophile Angriff des katalytisch essentiellen Cysteins führt zur Bildung eines negativ geladenen tetraedrischen Intermediates. Nach Freisetzung von Ammoniak wird ein weiteres tetraedrisches Intermediates gebildet, dessen Kollabieren zum Ablösen von Glutamat führt.

Zusammenfassend ist festzuhalten, daß die ins Cluster gruppierten Enzyme eine gemeinsame Teilreaktion katalysieren, bei der Glutamin hydrolysiert und die Aminogruppe in Form von nascentem Ammoniak auf ein Akzeptorsubstrat übertragen wird, wobei eine neue Kohlenstoff-Stickstoff-Gruppe entsteht.

3.6.3.3 Adenylat-bildende Enzyme

Tabelle 3-16 zeigt Sequenzzusammensetzung und Sequenzzahlen eines weiteren zur Betrachtung ausgewählten Clusters.

Tabelle 3-16: Übersicht über die Sequenzzusammensetzung des Clusters

EC-Klasse	Enzym	Sequenzzahl			
		Cluster			Datensatz
		10^{-10}	10^{-5}	10^{-2}	
1.2.1.31	L-Aminoadipat-semialdehyd Dehydrogenase	6	6	6	7
1.13.12.7	*Photinus*-luciferin 4-Monooxygenase (ATP-hydrolysierend)	10	10	10	10
2.7.7.58	(2,3-Dihydroxybenzoyl)adenylat Synthase	3	3	3	3
5.1.1.11	Phenylalanin Racemase (ATP-hydrolysierend)	4	4	4	4
6.1.1.13	D-Alanin-poly(phosphoribitol) Ligase	17	17	17	17
6.2.1.1	Acetat-CoA Ligase	97	97	97	97
6.2.1.3.	langkettige Fettsäure-CoA Ligase	88	88	88	88
6.2.1.12	4-Cumarat-CoA Ligase	31	32	33	33
6.2.1.17	Propionat-CoA Ligase	7	7	7	7
6.2.1.26	o-Succinylbenzoat-CoA Ligase	19	19	19	19
6.2.1.30	Phenylacetat-CoA Ligase	0	5	6	7
6.2.1.32	Anthranilat-CoA Ligase	2	2	2	3
6.3.2.26	N-(5-Amino-5-carboxypentanoyl)-L-cysteinyl-D-valin Synthase	5	5	5	5

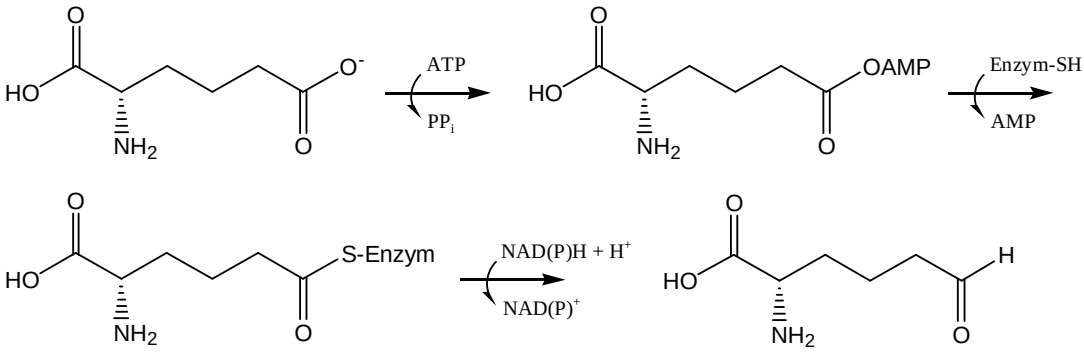
Das dargestellte Cluster enthält je nach Grenzwert zwischen 293 und 301 Sequenzen, die 12 bzw. 13 unterschiedlichen EC-Klassen angehören. Zwei der Enzymklassen gehören zur Hauptklasse der Oxidoreduktasen, je eine zu den Transferasen und Isomerasen und die übrigen zur Gruppe der Ligasen. Die bei den verschiedenen Grenzwerten nur geringe Änderung in der Zusammensetzung des Clusters deutet daraufhin, daß es sich um eine relativ klar definierte Sequenzfamilie handelt.

Tabelle 3-17: Übersicht über die von den Enzymen katalysierten Reaktionen

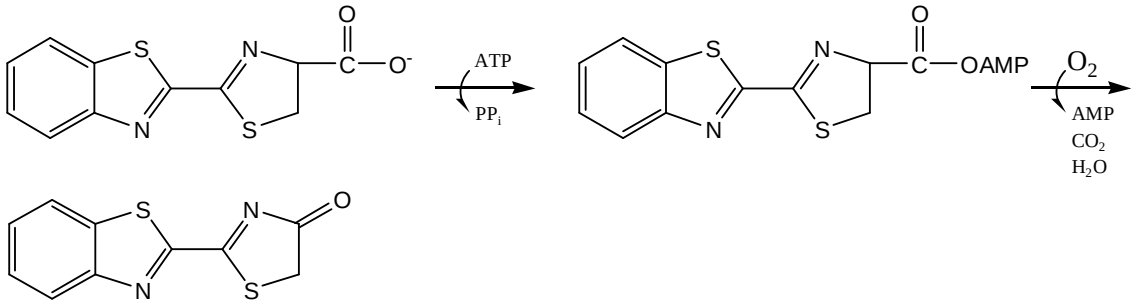
EC-Klasse	Enzym	katalysierte Reaktion
1.2.1.31	L-Aminoadipat-semialdehyd Dehydrogenase	L-2-aminoadipat-6-semialdehyd + $NAD(P)^+$ + H_2O = L-2-aminoadipat + NAD(P)H + H^+
1.13.12.7	*Photinus*-luciferin 4-Monooxygenase (ATP-hydrolysierend)	*Photinus* Luciferin + O_2 + ATP = oxidiertes *Photinus* Luciferin + CO_2 + AMP + Diphosphat + hη
2.7.7.58	(2,3-Dihydroxybenzoyl)adenylat Synthase	ATP + 2,3-Dihydroxybenzoat = Diphosphat + (2,3-Dihydroxybenzoyl)adenylat
5.1.1.11	Phenylalanin Racemase (ATP-hydrolysierend)	ATP + L-Phenylalanin + H_2O = AMP + Diphosphat + D-Phenylalanine
6.1.1.13	D-Alanin-poly(phosphoribitol) Ligase	ATP + D-Alanin + Poly(ribitolphosphat) = AMP + Diphosphat + O-D-Alanyl-poly(ribitol-phosphat)
6.2.1.1	Acetat-CoA Ligase	ATP + Acetat + CoA = AMP + Diphosphat + Acetyl-CoA
6.2.1.3.	langkettige Fettsäure-CoA Ligase	ATP + eine langkettige Carbonsäure + CoA = AMP + Diphosphat + ein Acyl-CoA
6.2.1.12	4-Cumarat-CoA Ligase	ATP + 4-Cumarat + CoA = AMP + Diphosphat + 4-Coumaroyl-CoA
6.2.1.17	Propionat-CoA Ligase	ATP + Propanoat + CoA = AMP + Diphosphat + Propanoyl-CoA
6.2.1.26	o-Succinylbenzoat-CoA Ligase	ATP + 2-Succinylbenzoat + CoA = AMP + Diphosphat + 2-Succinylbenzoyl-CoA
6.2.1.30	Phenylacetat-CoA Ligase	ATP + Phenylacetat + CoA = AMP + Diphosphat + Phenylacetyl-CoA
6.2.1.32	Anthranilat-CoA Ligase	ATP + Anthranilat + CoA = AMP + Diphosphat + Anthranilyl-CoA
6.3.2.26	N-(5-Amino-5-carboxypentanoyl)-L-cysteinyl-D-valin Synthase	L-2-Aminohexanedioat + L-Cystein + L-Valin + 3 ATP + H_2O = N-[L-5-Amino-5-carboxypentanoyl]-L-cysteinyl-D-valine + 3 AMP + 3 Diphosphat

Tabelle 3-17 zeigt, daß die gruppierten Enzyme zum Teil sehr unterschiedliche Reaktionen katalysieren. Die *Photinus*-luciferin 4-Monooxygenase katalysiert beispielsweise eine Redoxreaktion, bei der Luciferin in Gegenwart von ATP und Sauerstoff oxidiert und Licht erzeugt wird. Die (2,3-Dihydroxybenzoyl)adenylat Synthase katalysiert hingegen einen Nucleotidylgruppentransfer, die Phenylalanin Racemase eine Racemisierungsreaktion und die N-(5-Amino-5-carboxypentanoyl)-L-cysteinyl-D-valin Synthase die Bildung von Peptidbindungen. Eines haben jedoch alle diese Reaktionen gemeinsam. Der erste Schritt besteht in der Aktivierung des Substrats durch Bildung eines Adenylats:

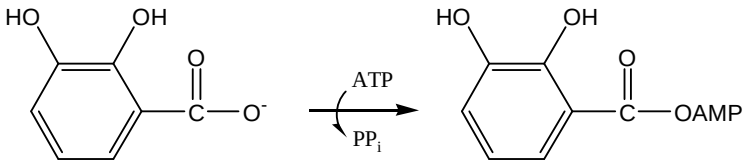
L-Aminoadipat-semialdehyd Dehydrogenase:



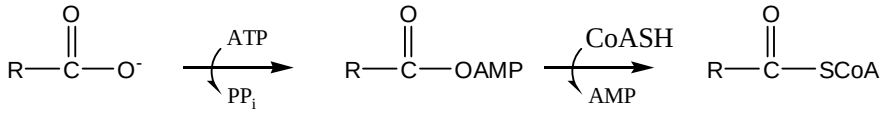
Photinus-luciferin 4-Monooxygenase:



(2,3-Dihydroxybenzoyl)adenylat Synthase:



Acyl:CoA Ligasen:



Diese funktionelle Gemeinsamkeit spiegelt sich auch in einem gemeinsamen Sequenzmotiv wider, das gewisse Ähnlichkeiten zur ATP-/GTP-bindenden „P-loop"-Domäne [119] und zu den ATP-bindenden Domänen der Proteinkinasen aufweist [120] und bei dem es sich laut PROSITE wahrscheinlich um ein AMP-Bindungsmotiv handelt (AMP_BINDING (PS00455)). Neben diesem Sequenzmotiv gibt es zudem noch weitere konservierte Sequenzpositionen, die in der Primärstruktur jedoch zum Teil weit auseinander liegen. Um deren mögliche Beteiligung an der Reaktion zu eruieren, wurde ihre Lage in der Tertiärstruktur betrachtet.

Die Struktur der *Photinus*-luciferin 4-Monooxygenase besteht aus einer großen N- und einer kleineren C-terminalen Strukturdomäne, die durch einen tiefen Spalt voneinander getrennt sind, von dem angenommen wird, daß er das katalytische Zentrum enthält [121] (vgl. Abbildung 3-19). Die Tatsache, daß vier der fünf konservierten Sequenzpositionen ebenfalls dort lokalisiert sind (vgl. Abbildung 3-20), bekräftigt die Vermutung, daß sie an der Bindung bzw. Bildung des Acyl-Adenylats-Intermediats beteiligt sind.

Abbildung 3-19: Proteinstruktur der *Photinus*-luciferin 4-Monooxygenase (PDB-ID 2D1S). Das Enzym besteht aus zwei durch einen Spalt voneinander getrennten Strukturdomänen.

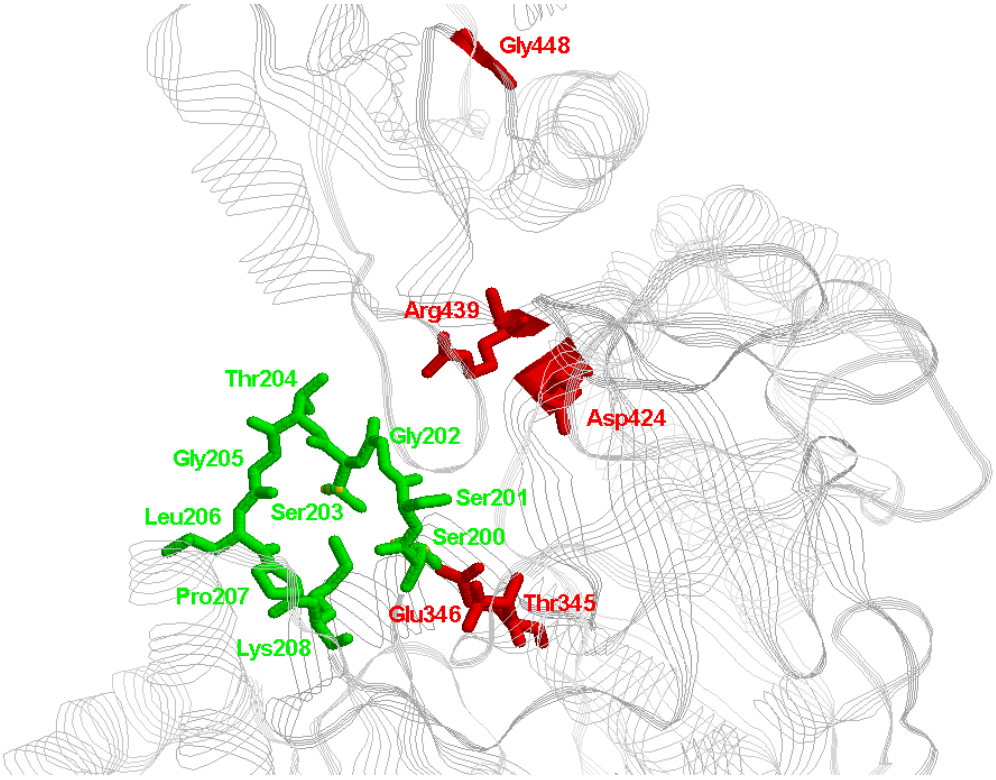


Abbildung 3-20: Schematische Darstellung des aktiven Zentrums der *Photinus*-luciferin 4-Monooxygenase (PDB-ID 2D1S). Das aktive Zentrum setzt sich aus den Bereichen zweier Strukturdomänen zusammen. Die konservierten Aminosäurepositionen sind farbig hervorgehoben und entsprechend ihrer Position gekennzeichnet. Grün: Aminosäuren des AMP-Bindungsmotiv, rot: weiter konservierte Sequenzpositionen.

Die Region des AMP-Bindungsmotiv (SSGSTGLPK) – in Abbildung 3-20 grün dargestellt - ist auf einer flexiblen Schleife (*Loop*) lokalisiert. Vier der übrigen fünf konservierten Sequenzpositionen (Thr345, Glu346, Asp424 und Arg439) befinden sich in unmittelbarer Nachbarschaft dem *Loop* gegenüber. Somit ist eine Beteiligung dieser Aminosäuren am katalytischen Mechanismus nicht nur möglich sondern in Anbetracht ihrer hohen Konservierung innerhalb der verschiedenen Enzyme auch wahrscheinlich.

Zusammenfassend ist somit festzuhalten, daß alle ins Cluster gruppierten Enzyme ein gemeinsames AMP-Bindungsmotiv sowie weitere, vermutlich an der Reaktion beteiligte konservierte Aminosäurepositionen besitzen und in einer gemeinsamen Teilreaktion eine ATP-abhängigen kovalent Bindung von AMP an das jeweilige Substrat katalysieren.

3.6.3.4 Glykosylhydrolasen

Tabelle 3-18 zeigt die Sequenzzusammensetzung sowie die Sequenzzahlen eines weiteren zur Betrachtung ausgewählten Clusters.

Tabelle 3-18: Übersicht über die Sequenzzusammensetzung des Clusters

EC-Klasse	Enzym	Sequenzzahl			
		Cluster			Datensatz
		10^{-10}	10^{-5}	10^{-2}	
2.4.1.4	Amylosucrase	1	1	1	1
2.4.1.5	Dextransucrase	1	11	14	16
2.4.1.7	Sucrose Phosphorylase	8	8	8	8
2.4.1.18	1,4-α-Glucan verzweigendes Enzym	96	96	96	96
2.4.1.19	Cyclomaltodextrin Glucanotransferase	21	21	21	21
2.4.1.25	4-α-Glucanotransferase	3	3	5	32
2.4.1.81	Flavone 7-*O*-β-Glucosyltransferase	3	3	3	3
2.4.1.140	Alternansucrase	0	1	1	1
2.4.1.183	α-1,3-Glucan Synthase	4	4	4	5
3.2.1.1	α-Amylase	182	184	184	193
3.2.1.10	oligo-1,6-Glucosidase	15	15	15	19
3.2.1.20	α-Glucosidase	20	20	20	42
3.2.1.33	Amylo-1,6-Glucosidase	0	0	2	2
3.2.1.41	Pullulanase	16	16	16	16
3.2.1.54	Cyclomaltodextrinase	9	9	9	9
3.2.1.60	Glucan 1,4-α-Maltotetraohydrolase	2	2	2	2
3.2.1.68	Isoamylase	8	8	9	10
3.2.1.70	Glucan 1,6-α-Glucosidase	7	7	7	7
3.2.1.93	α,α-Phosphotrehalase	9	9	9	9
3.2.1.98	Glucan 1,4-α-Maltohexaosidase	3	3	3	3
3.2.1.133	Glucan 1,4-α-Maltohydrolase	4	4	4	4
3.2.1.135	Neopullulanase	6	6	6	6
3.2.1.141	4-α-D-{(1→4)-α-D-Glucano} trehalose Trehalohydrolase	6	6	6	6
5.4.99.15	(1→4)-α-D-Glucan 1-α-D-Glucosylmutase	0	2	2	2
5.4.99.16	Maltose α-D-Glucosyltransferase	2	2	2	2

Das dargestellte Cluster enthält abhängig vom Grenzwert zwischen 424 und 446 verschiedene Sequenzen. Diese sind 22 bis 25 unterschiedlichen Enzymklassen zugeordnet, die drei verschiedenen Hauptklassen angehören. Wie auch schon im Beispiel des größten generierten Clusters (vgl. Abschnitt 3.6.2), konnten auch für dieses Cluster keine vollständig konservierten Sequenzpositionen im multiplen Alignment ermittelt werden. Die Recherche ergab jedoch, daß dennoch eine evolutionäre wie auch mechanistische Verwandtschaft zwischen diesen Enzymen besteht.

Die Datenbank CAZy (Carbohydrate-Active enZymes) [122] beschreibt Familien von strukturell verwandten katalytischen und Kohlenhydrat-bindenden Enzymdomänen, die glykosidische Bindungen spalten, knüpfen oder modifizieren. Die dortige Einteilung beruht zum einen auf der sequentiellen und zum anderen auf der strukturellen Ähnlichkeit der Enzyme. Demnach können bei den Glykosylhydrolasen 106 Sequenz- (GH-1 bis GH-106) und 14 Strukturfamilien (GH-A bis GH-N) unterschieden werden. Ein Vergleich der Einteilung mit dem anhand der eigenen Methode erhaltenen Cluster ergab, daß die zusammengefaßten Enzyme lediglich zwei der 106 Sequenzfamilien angehören (GH-13 und GH-70). Diese beiden Sequenzfamilien bilden überdies mit einer weiteren (GH-77) eine eigenständige Strukturfamilie. Diese Familie ist durch ein $(\beta\alpha)_8$-Faß charakterisiert, welches auch das aktive Zentrum enthält [123].

Glykosylhydrolasen katalysieren die hydrolytische Spaltung glykosidischer Bindungen zwischen zwei und mehr Kohlenhydraten bzw. zwischen einem Kohlenhydrat und einer Nicht-Kohlenhydrat-Einheit. Die Hydrolyse der glykosidischen Bindung erfolgt mittels einer allgemeinen Säurekatalyse und erfordert einen Protonendonor und ein Nukleophil/eine Base [124]. Die Hydrolyse kann über einen von zwei grundlegenden Mechanismen verlaufen, die in einer Inversion (Umkehrung) bzw. Retention (Beibehaltung) der anomeren Konfiguration resultieren [125, 126].

Die invertierenden Enzyme arbeiten mittels eines einstufigen Mechanismus, bei dem ein Basen-unterstützter nukleophiler Angriff eines Wassermoleküls die Konfiguration am anomeren Zentrum des Substrates umkehrt (invertiert) (vgl. Abbildung 3-21a). Bei den retendierenden Enzymen erfolgt die Reaktion über einen zweistufigen Mechanismus, der unter Beteiligung zweier katalytischer Carboxylgruppen verläuft (vgl. Abbildung 3-21b). Im ersten Schritt (Gly-

kosylierung) liefert eine Carboxylgruppe ein Proton, während zeitgleich ein nukleophiler Angriff des zweiten Carboxylats zur Bildung eines Glykosyl-Enzym-Zwischenprodukts führt. Im zweiten Schritt (Deglykosylierung) fungiert die erste Carboxylgruppe als Base und aktiviert ein Nukleophil (eine Wassermolekül im Falle der Hydrolyse und einen Alkohol im Falle der Transglykosylierung), welches das Glykosyl-Enzym hydrolysiert. Infolgedessen entsteht ein Produkt mit derselben anomeren Konfiguration wie das Substrat.

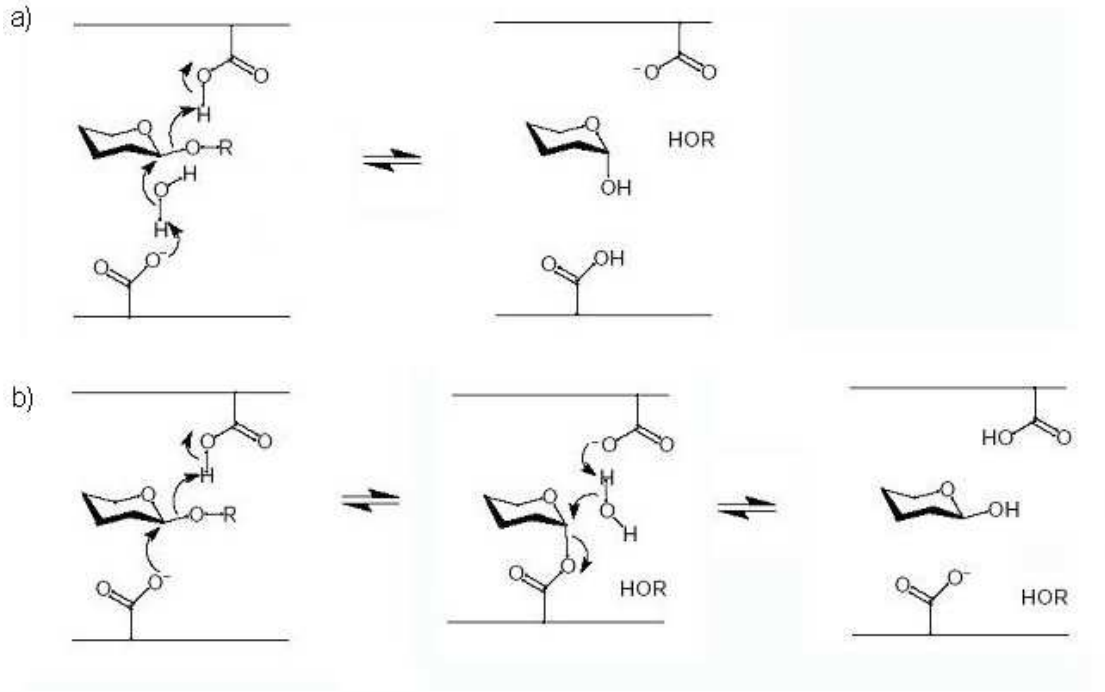


Abbildung 3-21: Mechanismen der enzymatischen Hydrolyse der glykosidischen Bindungen bei a) invertierenden und b) retendierenden Enzymen.

Quelle: Carbohydrate-Active enZYmes (http://afmb.cnrs-mrs.fr/CAZY/)

Beide im Cluster befindlichen Sequenzfamilien gehören zur Gruppe der retendierenden Enzymen. Die im Cluster zusammengefaßten Enzyme besitzen somit nicht nur einen gemeinsamen evolutionären Ursprung sondern verwenden überdies auch denselben enzymatischen Reaktionsmechanismus zur Realisierung ihrer unterschiedlichen Funktionen. Somit konnte die von CAZy vorgenommene Klassifizierung mittels der eigenen Methode nicht nur reproduziert sondern überdies auch zwei Sequenzfamilien miteinander assoziiert werden, die anhand der für CAZy verwendeten Methode erst auf Strukturebene miteinander korreliert wurden.

3.6.3.5 Fumarat Lyasen

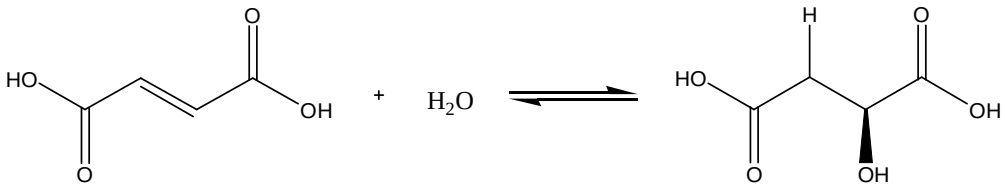
In der nachfolgende Tabelle 3-19 ist die Sequenzzusammensetzung des betrachteten Clusters sowie die Zahl der im Cluster und der im gesamten Datensatz enthaltenen Sequenzen angegeben.

Tabelle 3-19: Übersicht über die Sequenzzusammensetzung des Clusters

EC-Klasse	Enzym	Sequenzzahl			
		Cluster			Datensatz
		10^{-10}	10^{-5}	10^{-2}	
4.2.1.2	Fumarat-Hydratase	87	87	87	105
4.3.1.1	Aspartat-Ammoniak-Lyase	20	20	20	20
4.3.2.1	Argininosuccinat-Lyase	120	120	120	120
4.3.2.2	Adenylosuccinat-Lyase	60	60	60	60
5.5.1.2	3-Carboxy-cis,cis-muconat Cycloisomerase	8	8	8	8

Das dargestellte Cluster enthält 295 Sequenzen, die fünf unterschiedlichen Enzymklassen angehören. Vier von diesen zählen zur Hauptklasse der Lyasen, eine dagegen zur Hauptklasse der Isomerasen. Die gleichbleibende Größe des Clusters auch bei höheren Grenzwerten indiziert, daß die Sequenzen innerhalb des generierten Clusters eine klar definierte Sequenzfamilie bilden.

Das Enzym Fumarat-Hydratase (Fumarase) katalysiert die reversible, stereospezifische trans-Addition von Wasser (OH^-, H^+) an die Doppelbindung von Fumarat. Da die OH-Gruppe ausschließlich an eine Seite der Doppelbindung addiert wird, entsteht bei dieser Hydrierung nur das L-Isomer des Malats [127]:



Fumarat (*S*)-Malat (L-Malat)

Abbildung 3-22: Schematische Darstellung der Fumarase-Reaktion.

In der Literatur werden zwei verschiedene Klassen der Fumarasen unterschieden [128]. Bei den Fumarasen der Klasse I handelt es sich um thermolabile Dimere, die ausschließlich bei Prokaryoten auftreten. Bei den Klasse II Fumarasen handelt es sich um thermostabile Tetramere, die sowohl in Prokaryoten als auch Eukaryoten vorkommen. Die Sequenzen der beiden Klassen sind nicht eng miteinander verwandt [128]. Dies spiegelt sich auch in der Zusammensetzung des Clusters wider. Während bei den übrigen vier Enzymklassen jeweils alle im Datensatz enthalten Sequenzen ins Cluster gruppiert wurden, sind von den 105 Fumarase-Sequenzen lediglich 87 im Cluster enthalten. Eine Überprüfung ergab, daß es sich bei diesen ausnahmslos um Klasse II Fumarasen handelt. Die 18 zur Klasse I gehörenden Sequenzen bilden hingegen ein eigenständiges Cluster.

Die Aspartat-Ammoniak-Lyase (Aspartase) katalysiert die reversible Umsetzung von Aspartat zu Fumarat und Ammoniak. Diese Reaktion ist analog zur Rückreaktion der Fumarase, außer das Ammoniak anstelle von Wasser an der trans-Eliminierungsreaktion beteiligt ist [129].

Das Enzym Argininosuccinat-Lyase (Argininosuccinase) ist ubiquitär in allen Organismen vorhanden [130]. Es ist in Prokaryoten an der Biosynthese von Arginin und in Eukaryoten am Harnstoffzyklus beteiligt und katalysiert die Bildung von Arginin und Fumarat aus Argininosuccinat.

Die Adenylosuccinat-Lyase (Adenylosuccinase) katalysiert die Bildung von 5-Amino-1-(5-phospho-D-ribosyl)imidazol-4-carboxamid und Fumarat aus (*S*)-2-[5-Amino-1-(5-phospho-D-ribosyl)imidazol-4-carboxamido]succinat. Das Enzym kann zudem die Bildung von Fumarat und AMP aus Adenylosuccinat katalysieren [131]. Diese Reaktionen lassen sich schematisch wie folgt zusammenfassen:

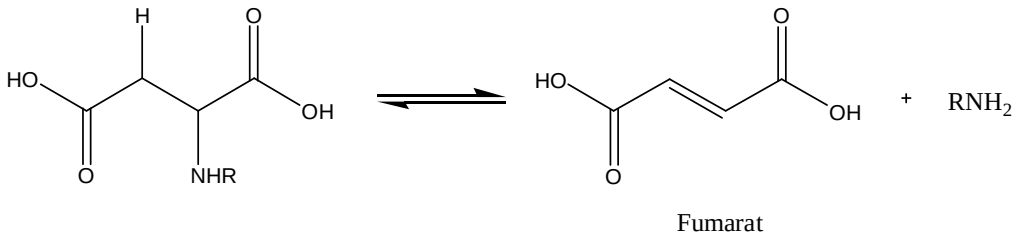
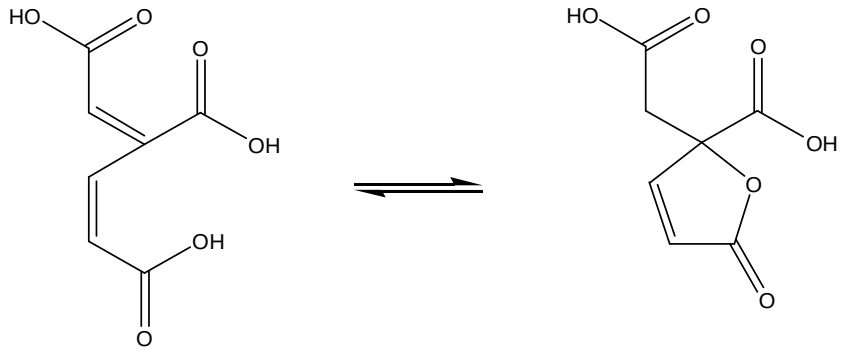


Abbildung 3-23: Schematische Darstellung der Aspartase-, Arginosuccinase- und Adenylosuccinase-Reaktion.

Das Enzym 3-Carboxy-cis,cis-muconat-Cycloisomerase (3-Carboxymuconat lactonisierendes Enzym) ist am Katabolismus aromatischer Säuren beteiligt und katalysiert in einer Isomerisierungsreaktion die Lactonisierung von cis,cis-Butadien-1,2,4-tricarboxylat, was zur Bildung von 2-Carboxy-2,5-dihydro-5-oxofuran-2-acetat führt [132]:



cis,cis-Butadien-1,2,4-tricarboxylat 2-Carboxy-2,5-dihydro-5-oxofuran-2-acetat

Abbildung 3-24: Schematische Darstellung der 3-Carboxy-cis,cis-muconat-Cycloisomerase-Reaktion.

Bei den von den Lyasen katalysierten Umsetzungen handelt es sich um Eliminierungsreaktionen. Eliminierungsreaktionen führen zur Bildung einer Doppelbindung zwischen zwei zuvor einfach gebundenen, gesättigten Zentren. Eliminiert werden können z. B. Wasser, Ammoniak, Alkohole (ROH) oder primäre Amine (RNH_2).

Bindungsspaltung und –bildung können bei dieser Reaktion über einen von drei unterschiedlichen Mechanismen verlaufen (vgl. Abbildung 3-25):

1. konzertiert,
2. stufenweise, wobei durch die Spaltung der C-O-Bindung zunächst ein Carbokation gebildet wird, oder
3. stufenweise, wobei durch die Spaltung der C-H-Bindung zunächst ein Carbanion gebildet wird.

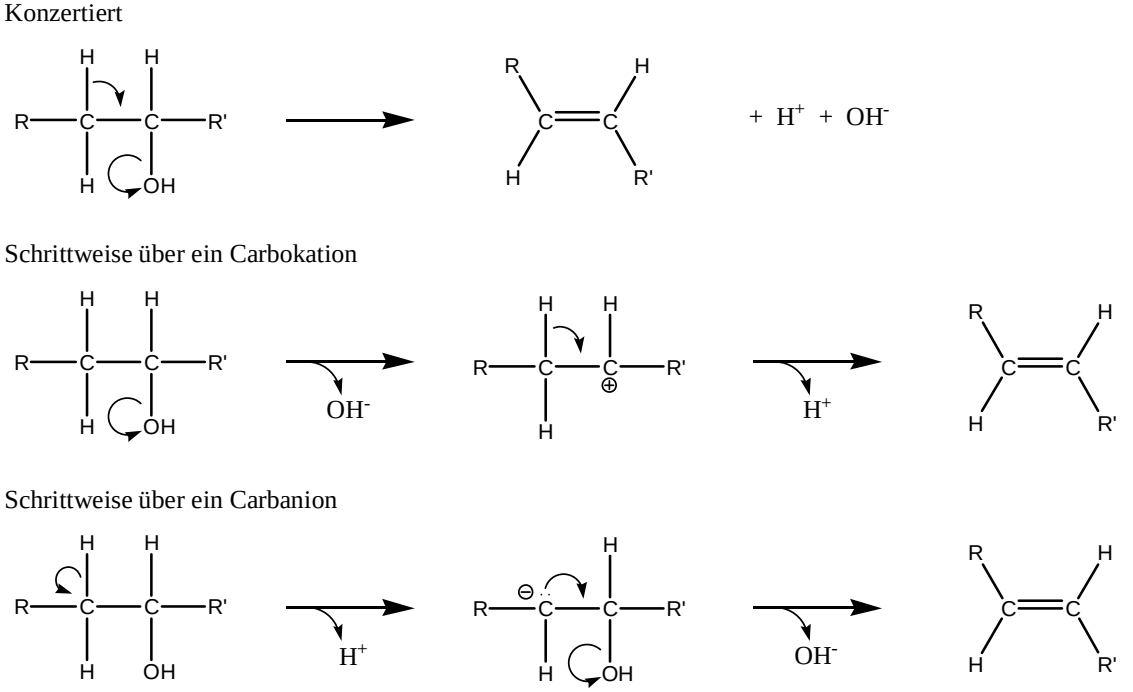


Abbildung 3-25: Mögliche Mechanismen für eine Eliminierungsreaktion am Beispiel der Dehydrierung. Die Reaktionen können konzertiert, schrittweise über ein Carbokation als Zwischenprodukt oder schrittweise über ein Carbanion als Zwischenprodukt verlaufen [1].

Biochemische Isomerisierungsreaktionen verlaufen unter intramolekularer Verschiebung eines Wasserstoffatoms, wodurch sich die Lage einer Doppelbindung verschiebt. Dabei wird ein Proton von einem C-Atom abgespalten und an ein anderes angelagert. Folglich kann die von der 3-Carboxy-cis,cis-muconat Cycloisomerase katalysierte Isomerisierungsreaktion als eine intramolekulare Eliminierungs- und Additionsreaktion aufgefaßt werden.

Aufgrund ihrer essentiellen Rolle beim Citrat-Zyklus gehört die Fumarase zu den am besten untersuchten Enzymen im Cluster. Experimente, die zeigen sollten, ob die Fumarase-Reaktion über einen Carbanion-Mechanismus (1. Schritt: OH^--Addition) oder einen Carbokation-Mechanismus (1. Schritt: H^+-Addition) verläuft, lieferten einander widersprechende Ergebnisse (vgl. Abbildung 3-26). Die Ergebnisse, die für einen Carbokation-Mechanismus sprechen, wurden durch eine Untersuchung der Dehydratation von L-Malat (die Umkehrung der Fumarase-Reaktion) in $H_2{}^{18}O$ erhalten. [^{18}O]-Malat erscheint rascher im Reaktionsgemisch, als es der Fall wäre, wenn ^{18}O über eine Rückreaktion des neugebildeten Fumarats eingebaut würde. Dies spricht für die rasche Bildung eines

an C2 positiv geladenen Zwischenprodukts (aus dem heraus OH^- mit $^{18}OH^-$ austauschen könnte), gefolgt von einer langsamen Wasserstoffabspaltung an C3 [133].

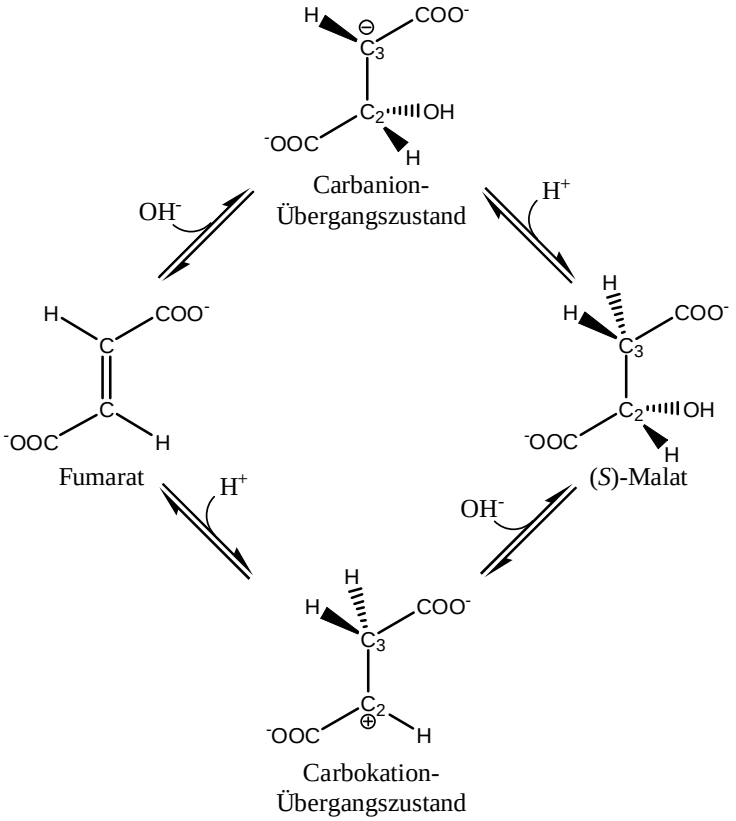
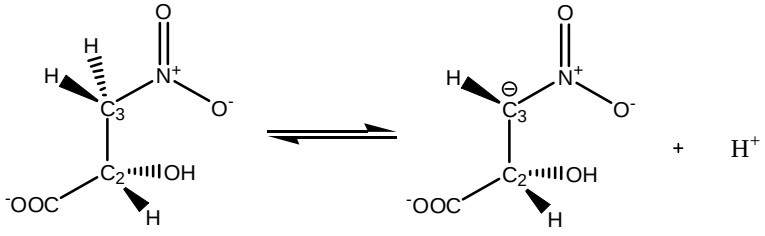


Abbildung 3-26: Mögliche Mechanismen für die durch Fumarase katalysierte Hydratation von Fumarat [1].

Andere Beobachtungen deuten jedoch darauf hin, daß die Reaktion über die Bildung eines an C3 negativ geladenen Zwischenprodukts verläuft. Porter *et al*. (1980) synthetisierte (*S*)-2-hydroxy-3-nitropropionat, das (*S*)-Malat sterisch ähnelt [134]:



(*S*)-2-hydroxy-3-nitropropionat

Aufgrund des elektronenziehenden Charakters der Nitrogruppe sind die C3-Protonen relativ sauer (pK ~10). Das resultierende Anion ist ein Analogon des postulierten C3-carbanionischen Übergangszustandes der Fumarase-Reaktion, nicht aber des C2-carbokationischen Übergangszustandes (vgl. Abbildung 3-26). Dieses Anion hat sich als ein äußerst wirksamer Hemmstoff der Fumarase erwiesen. Es hat eine vielfach höhere Bindungsaffinität zu dem Enzym als (*S*)-Malat [134]. Die Befunde beider Untersuchungen stehen somit im Widerspruch zueinander. Tatsächlich ergibt eine andere Interpretation des ^{18}O-Austauschexperimentes, daß es auch mit einem Carbanion-Mechanismus vereinbar ist, wenn OH^- sehr rasch vom Enzym freigesetzt wird und die Freisetzung der anderen Reaktionsprodukte der geschwindigkeitsbestimmende Schritt ist. Somit kann $^{18}OH^-$ mit OH^- austauschen, und [^{18}O]-Malat kann mit einer höheren Geschwindigkeit als der Gesamtgeschwindigkeit der Fumarase-Reaktion gebildet werden [135].

Die Umsetzung der ionisierten Zwischenform des Enzym-Fumarat-Komplexes zum Enzym-Malat-Komplex erfolgt vermutlich unter Beteiligung ionisierbarer Gruppen im katalytischen Zentrum, von denen eine notwendigerweise als Base (Abstraktion des Protons) und die andere als Säure (Protonierung der OH-Gruppe) agiert [136]. Die Beobachtung, daß bei der Fumarase-Reaktion nur Monodeutero-L-Malat entsteht, wenn die Reaktion in D_2O ausgeführt wird, stimmt mit der Annahme überein, daß das Proton direkt von einer aciden Gruppe des enzymatischen Zentrums übertragen wird [137, 138]. Neben den Aminosäuren, die als Protonenakzeptor und –donor wirken, werden überdies noch weitere Residuen benötigt.

Fumarase besitzt eine recht hohe Substratspezifität. Dennoch ist neben den natürlichen Substraten Fumarat und L-Malat, mittlerweile eine Reihe weiterer Substrate der Fumarase bekannt, von denen aber keines eine biologische Bedeutung besitzt, da sie, wenn überhaupt, nur in sehr geringen Mengen im Gewebe vorkommen [139]. Die Eigenschaften dieser Substrate vermitteln aber einen Eindruck von den katalytischen Eigenschaften der Fumarase. Betrachtet man deren Struktur, so ergeben sich zwei grundsätzliche Vorraussetzungen für die Substrate der Fumarase [140]:

- Das Substrat muß zwei negative geladene Carboxylgruppen aufweisen.

- Die Hydroxylgruppe muß die gleiche Konfiguration wie in (*S*)-Malat besitzen.

Ersteres läßt vermuten, daß zwei Regionen des aktiven Zentrums eine positive Ladung tragen, die mit den negativ geladenen Carboxylgruppen interagieren [140]. Ferner kann angenommen werden, daß das geladene Zwischenprodukt einer schrittweise verlaufenden Reaktion durch eine entgegengesetzt geladene Gruppe des aktiven Zentrums stabilisiert wird [140]. Auf Grundlage all dieser Informationen kann für die Fumarase folgender Reaktionsmechanismus formuliert werden:

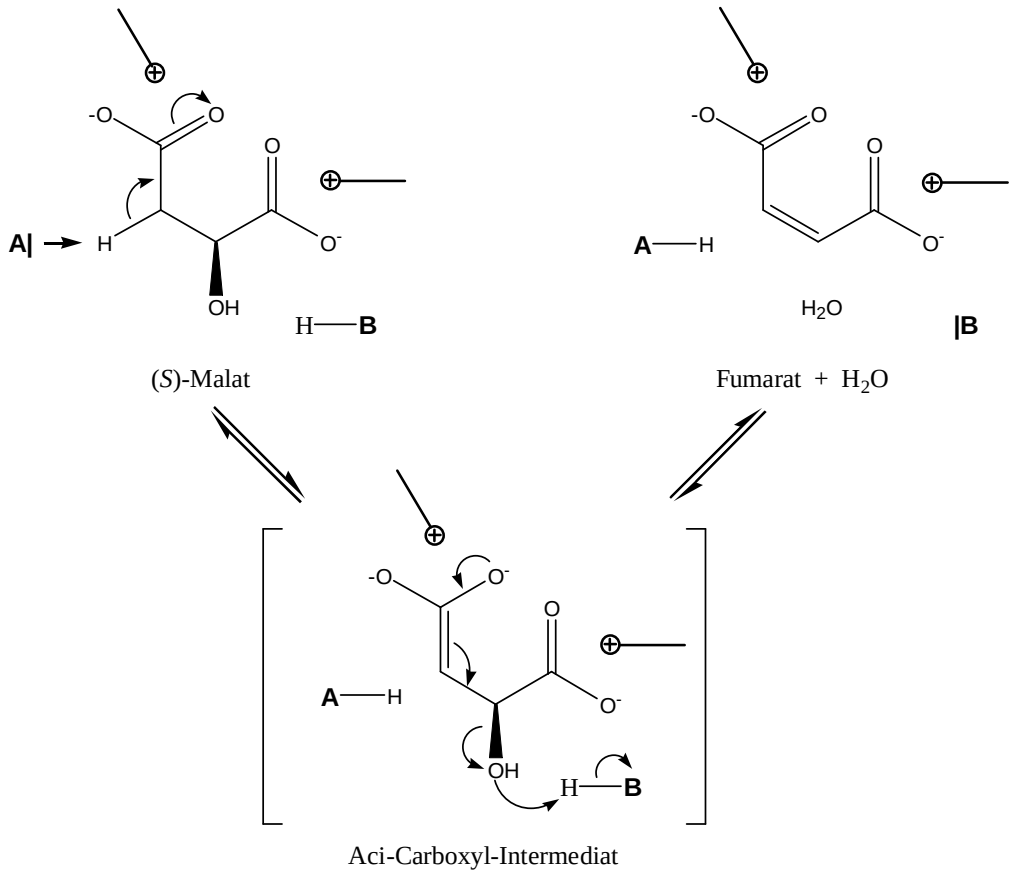


Abbildung 3-27: Schematische Darstellung des Reaktionsmechanismus der Fumarase (modifiziert nach Alberty *et al.* [141]). Das katalytische Zentrum wird durch vier nicht-identische Gruppen repräsentiert. Die positiv geladenen Gruppen interagieren mit den Carboxylatgruppen des Substrats, während A und B den beiden aciden Gruppen des katalytischen Zentrums entsprechen.

Eine basische Aminosäure abstrahiert ein Proton am C3-Atom unter Bildung eines Carbanions. Durch Elektronenverschiebung läßt sich eine der Aci-Struktur verwandte Carboxylstruktur als Übergangzustand formulieren. Die Hydroxylgruppe wird durch eine saure Aminosäure protoniert und die Doppelbindung des Fumarats generiert. Der formulierte Mechanismus entspricht formell einer E1cB-Eliminierungsreaktion und läßt sich auch auf die übrigen Fumarat-bildenden Lyasen im Cluster übertragen.

Über den Reaktionsmechanismus der 3-Carboxy-cis,cis-muconat-Cycloisomerase ist bisher nur wenig bekannt. Basierend auf dem Ergebnis zum chemischen Mechanismus der Fumarase kann für dieses Enzym aber analog folgender Reaktionsmechanismus postuliert werden:

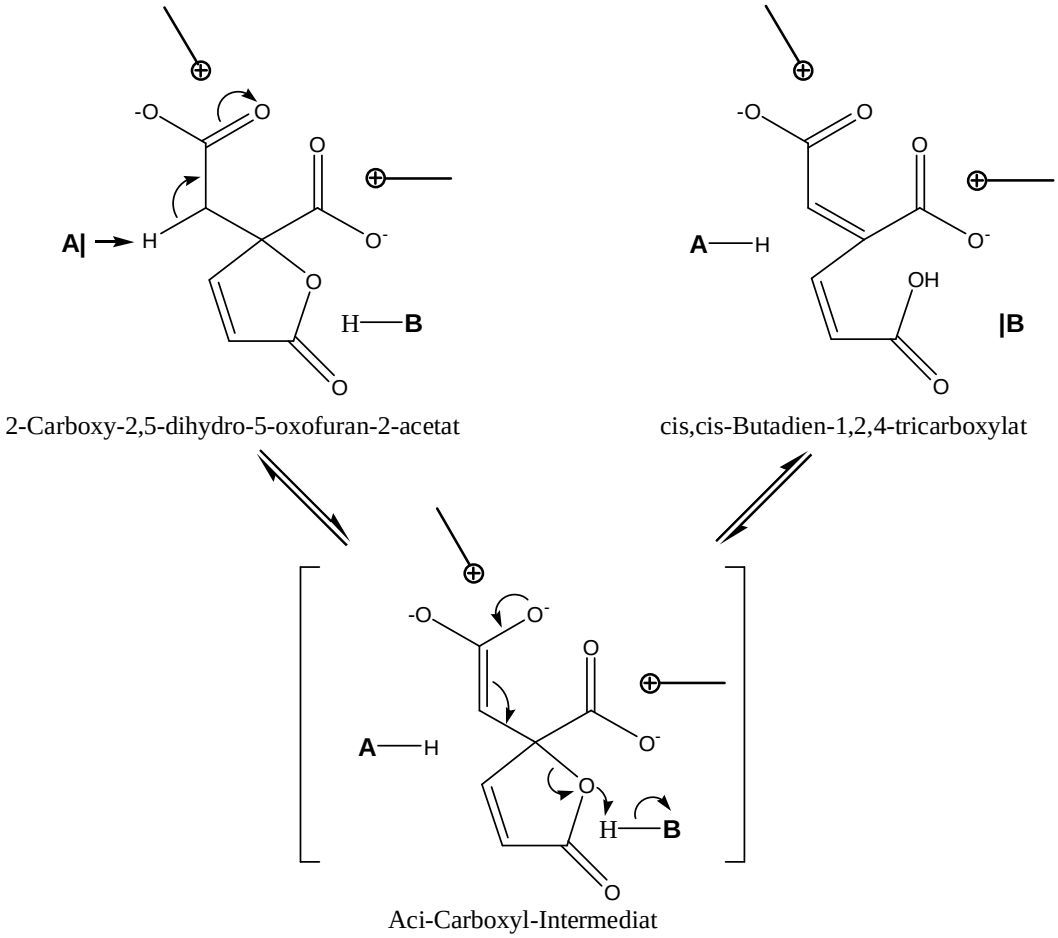


Abbildung 3-28: Schematische Darstellung des von Fumarase abgeleiteten Reaktionsmechanismus der 3-Carboxy-cis,cis-muconat-Cycloisomerase. Die positiv geladenen Gruppen interagieren mit den Carboxylatgruppen des Substrats, während A und B den beiden aciden Gruppen des katalytischen Zentrums entsprechen.

Es wurden mehrere Versuche unternommen, die Strukturen im aktiven Zentrum der Fumarase mittels Affinitätsmarkierungen zu identifizieren. Dazu wurde die Reaktion der Fumarase mit einigen Verbindungen untersucht, deren Struktur entweder der des Fumarats oder des L-Malats ähnelten und zudem eine Gruppe enthielten, die in der Lage war, mit den Seitenketten bestimmter Aminosäuren zu reagieren [142]. Es zeigte sich, daß Jodacetat eine 80-90%ige Inaktivierung der Fumarase bewirkt, ohne dabei die Proteinstruktur zu verändern. Aufgrund dieser Beobachtung wurde angenommen, daß das aktive Zentrum der Fumarase durch das Reagenz modifiziert worden war. Im Hydrolysat des Jodacetat-inaktivierten Enzyms konnten zwei modifizierte Aminosäuren identifiziert werden: 3-Carboxymethylhistidin und S-Carboxymethylmethionin. Gemäß dieser Studie scheinen somit Histidin sowie Methionin am katalytischen Mechanismus beteiligt zu sein. Es ist jedoch unklar, wie Methionin die Eigenschaften einer der geforderten Gruppen einnehmen soll. Dies schließt aber nicht aus, daß es auf eine andere Weise an der Reaktion beteiligt ist.

Die ergiebigsten Informationen über die an der Enzymkatalyse beteiligten Aminosäuren liegen für δ-Crystallin vor. δ-Crystallin ist das mengenmäßig wichtigste lösliche Protein in den Augenlinsen von Vögeln und Reptilien. Es entstand ursprünglich aus dem Argininosuccinat-Lyase Gen. Durch Duplikation entwickelten sich zwei Isoformen, das δ1- und δ2-Crystallin [143, 144]. Nur das δ2-Crystallin behielt während der Evolution die Argininosuccinat-Lyase Aktivität bei [145, 146]. Das δ1-Crystallin wandelte sich zu einem Strukturprotein in der Augenlinse und verlor in diesem Prozeß die enzymatische Funktion [147, 148]. δ1-Crystallin ist somit ein Beispiel für ein sogenanntes „*hijacked*“ Enzym [149]. Vergleiche der Aminosäuresequenzen von δ1- und δ2-Crystallin aus *Anasplatyrhynchos* belegen eine Sequenzidentität von 94% [143]. Die nur geringen Sequenzunterschiede zwischen der katalytisch aktiven und der inaktiven Form ermöglichen es, die an der Katalyse beteiligten Aminosäuren einzugrenzen.

Mutagenesestudien mit δ2-Crystallin implizieren, daß Histidin 162 zum aktiven Zentrum gehört (vgl. Abbildung 3-29). Es fungiert direkt oder indirekt über Aktivierung eines Wassermoleküls als Protonenakzeptor [151, 152]. In Argininosuccinat-Lyasen [153] und Adenylosuccinat-Lyasen [154] wird das entsprechende Histidin ebenfalls als Protonenakzeptor favorisiert. Weaver *et al*.

(1995) vermuteten für die Fumarasen ein Protonen-Relais zwischen dem äquivalenten Histidin 188 und der Glutaminsäure 331 [150]. Als potentielle Säure gilt in δ2-Crystallin das hochkonservierte Serin 283 [155, 156]. Simpson (1994) nahm an, daß in δ2-Crystallin das hochkonservierte Lysin 289 zur Stabilisierung des Carbanions beiträgt [149].

Wenn die beschriebenen Aminosäuren tatsächlich am katalytischen Mechanismus dieser Enzymfamilie beteiligt sind, dann sollten sie sich in entsprechender Position in allen Enzymen des Clusters nachweisen lassen. Um dies zu eruieren, wurde ein multiples Sequenzalignment aller 295 im Cluster enthaltenen Sequenzen erstellt. Die Primärsequenzen der Enzyme zeigen nur eine geringe Übereinstimmung. Dennoch können drei konservierte Bereiche definiert werden (C1 – C3; vgl. Abbildung 3-29), die höchstwahrscheinlich an der Generierung des aktiven Zentrums und der Substratbindung beteiligt sind [150]. Entgegen vorheriger Beobachtungen [155] ergibt sich anhand des eigenen Alignments, daß in den Bereichen C1 und C2 nur jeweils eine Aminosäureposition in allen Sequenzen konserviert ist. Der Bereich C3 enthält eine kurze konservierte Sequenz um ein zentrales Methionin, zu der auch ein Prosite-Motiv existiert (Kennummer: PS00163). Allerdings werden nicht alle im Cluster enthaltenen Sequenzen durch dieses Prosite-Motiv erfaßt (vgl. Abschnitt 3.7.1).

	C1	C2	C3
δ2-Crystallin P24058(4.3.2.1)	^{114}SRNDQVVT^{121}	^{159}GYT**H**LQKAQP^{168}	^{282}G**S**SLMPQ**K**KNPDSLE^{296}
P05042(4.2.1.2)	^{137}SSNDVFPT^{144}	^{185}GRT**H**LQDATP^{194}	^{317}G**S**SIMPG**K**VNPTQCE^{331}
P04422(4.3.1.1)	^{143}STNDAYPT^{150}	^{188}GRTQLQDAVP^{197}	^{320}G**S**SIMPA**K**VNPVVPE^{334}
Q9X0I0(4.3.2.2)	^{93}TSSDVLDT^{100}	^{138}GRT**H**GVHAEP^{147}	^{261}G**S**SAMPH**K**KNPITCE^{175}
P32427(5.5.1.2)	^{105}TSQDAMDT^{112}	^{150}GRTWLQHATP^{159}	^{275}G**S**STMPH**K**RNPVGAA^{289}
	:	:	** *. * *

Abbildung 3-29: Für die im Cluster enthaltenen Sequenzen lassen sich drei konservierten Bereiche ermitteln (C1-C3; nach Sampaleanu *et al.*, 2001). Die Darstellung zeigt die entsprechenden Bereiche für je einen Vertreter der im Cluster enthaltenen EC-Klassen. Die im Fettdruck dargestellten Aminosäuren sind vermutlich am katalytischen Mechanismus der Enzyme beteiligt (siehe Text). Die mit „*“ versehenen Aminosäuren sind in allen Sequenzen des Clusters konserviert, während „:“ auf konservativ substituierte und „.“ auf semikonservativ substituierte Aminosäurepositionen hinweisen.

Das multiple Alignment belegt, daß sowohl das als potentielle Säure angesehene Serin als auch das zur Stabilisierung des Carbanions benötigte Lysin in allen Enzymen des Clusters konserviert sind. Beide sind im Bereich C3 lokalisiert. Das als Protonenakzeptor favorisiert Histidin aus dem Bereich C2 ist hingegen nicht strikt konserviert. So ist es beispielsweise in Aspartasen vielfach durch Glutamin ersetzt (vgl. Abbildung 3-29). In der 3-Carboxy-cis,cis-muconat-Cycloisomerase sind die Variationen noch vielfältiger, so daß an dieser Stelle neben Histidin auch Arginin, Tryptophan und Leucin vorkommen können. Es erscheint daher fraglich, ob besagtes Histidin tatsächlich als Protonenakzeptor fungiert. Anhand des Alignments läßt sich jedoch zeigen, daß die dem Histidin direkt benachbarte Position konservativ substituiert ist und ausschließlich von Serin oder Threonin eingenommen werden kann. Gleiches gilt für die konservierte Position im Bereich C1. Für δ2-Crystallin wird angenommen, daß die Hydroxylgruppe des Threonins 161 zur korrekten Positionierung der Seitenkette des Lysins 289 als auch des Fumarat-Anteils des Substrates benötigt wird [157]. Über die Bedeutung des Serins 114 ist bislang scheinbar noch nichts bekannt. Um dessen mögliche Beteiligung am Enzymmechanismus zu eruieren, wurde seine Lage in der Proteinstruktur betrachtet.

Zu den Sequenzen aller fünf EC-Klassen stehen aufgeklärte Proteinstrukturen zur Verfügung. Die enzymatisch aktive Form der Enzyme besteht aus vier identischen Untereinheiten (vgl. Abbildung 3-30). Jedes Enzym besitzt vier aktive Zentren, die sich aus den Bereichen dreier Untereinheiten zusammensetzen [150][158][159]. Bedingungen, die zur Dissoziation der Untereinheiten führen, resultieren im Verlust der enzymatischen Aktivität [139].

Um die Ähnlichkeit zwischen den Strukturen der verschiedenen EC-Klassen zu quantifizieren, wurden mittels des Programms Protein3Dfit [88] Strukturalignments erstellt. Standen mehrere Strukturen zu einer Enzymklasse zur Verfügung, so wurde jeweils die mit der besten Auflösung zur Betrachtung ausgewählt. Zwei Faltungen werden als strukturell ähnlich angesehen, wenn sich mindestens 40% der korrespondierenden C_α-Atome innerhalb eines Abstandes von maximal 1,8Å zueinander befinden. Tabelle 3-20 enthält eine Übersicht der Ergebnisse.

Tabelle 3-20: Übersicht über die strukturelle Ähnlichkeit zwischen den Enzymen der verschiedenen EC-Klassen.

EC-Klasse	Swiss-Prot Nummer	PDB-Struktur	1FUR	1JSW	1K7W	1C3C	1RE5
4.2.1.2	**P05042**	**1FUR**	-				
4.3.1.1	**P04422**	**1JSW**	63,16%	-			
4.3.2.1	**P24058**	**1K7W**	52,56%	44,54%	-		
4.3.2.2	**Q9X0I0**	**1C3C**	53,54%	44,10%	59,43%	-	
5.5.1.2	**P32427**	**1RE5**	52,97%	39,50%	54,11%	67,22%	-

Die Strukturalignments belegen nochmals die evolutionäre Verwandtschaft der miteinander gruppierten Sequenzen und bestätigen so die mittels Sequenzähnlichkeit getroffene Einteilung der Enzyme. Lediglich das Ergebnis eines Vergleichs (1RE5/1JSW) rangiert geringfügig unterhalb des geforderten Grenzwertes von 40%. Ihre Ähnlichkeit zu den Strukturen der übrigen EC-Klassen belegt aber auch ihre Verwandtschaft.

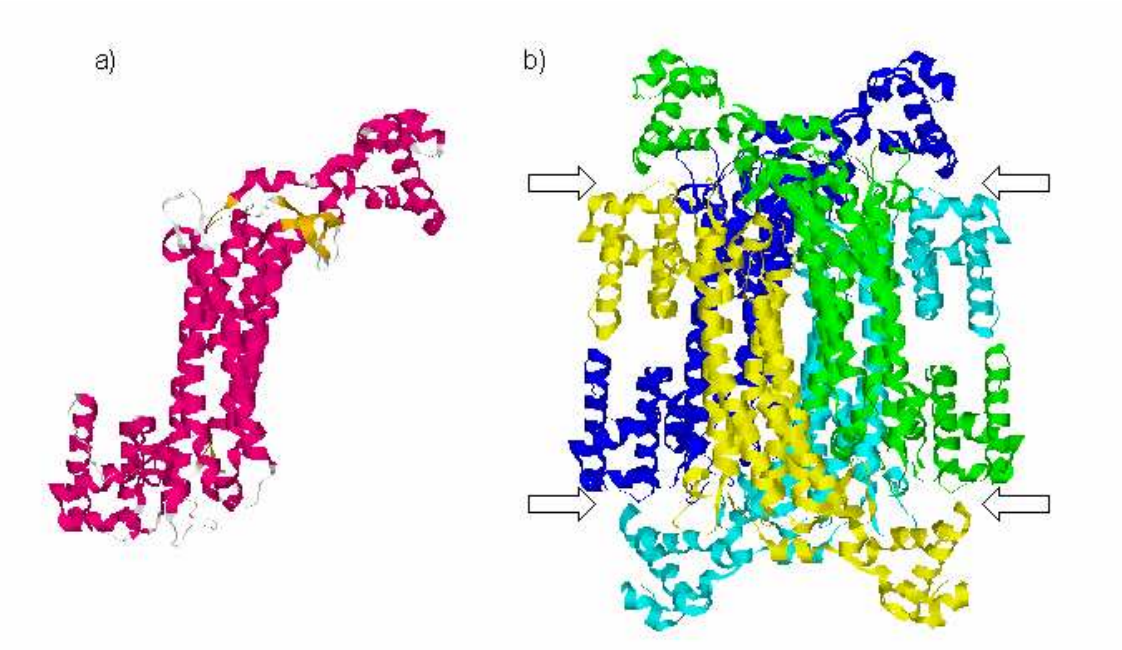


Abbildung 3-30: Proteinstruktur der Argininosuccinat-Lyase (PDB-ID 1K7W): a) Monomer, b) Tetramer. Das Enzym verfügt über eine homotetramere Struktur und besitzt vier aktive Zentren, deren Lage durch Pfeile markiert ist.

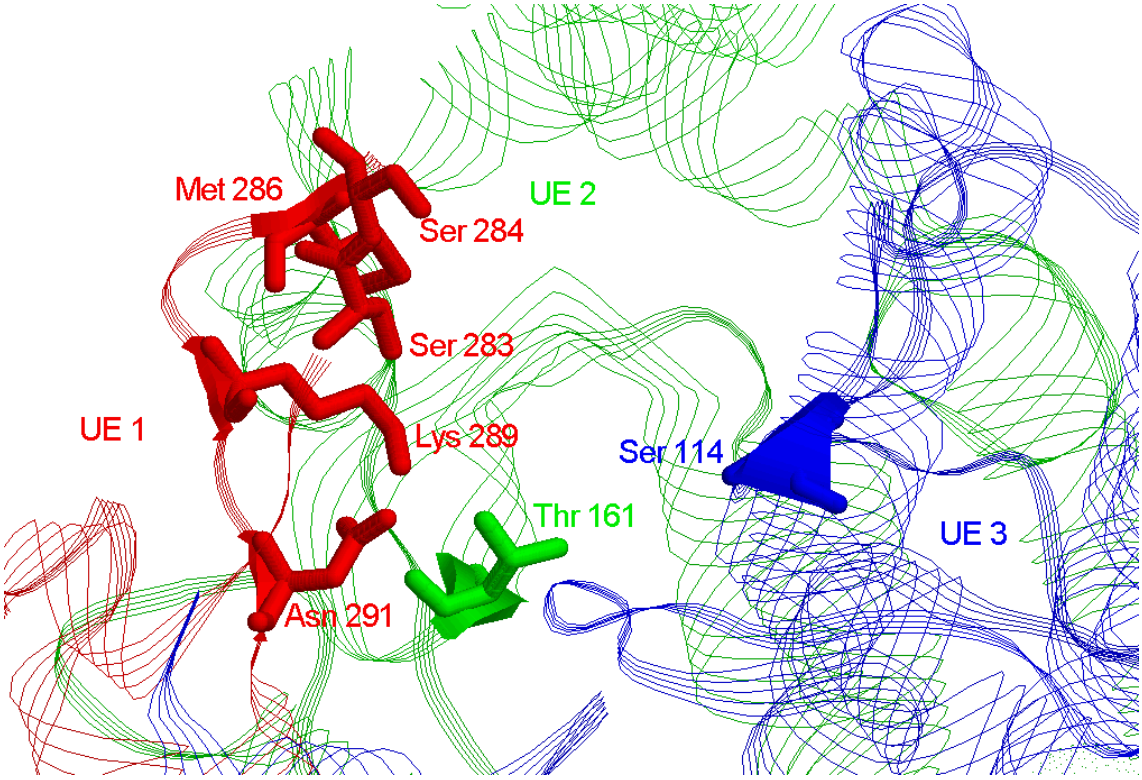


Abbildung 3-31: Schematische Darstellung eines der vier aktiven Zentren der Argininosuccinat-Lyase. Das aktive Zentrum setzt sich aus den Bereichen dreier Untereinheiten zusammen (UE 1-3). Die konservierten Aminosäurepositionen sind gekennzeichnet und entsprechend der Untereinheit, der sie angehören, koloriert.

Abbildung 3-31 zeigt, daß alle mittels des multiplen Sequenzalignments ermittelten Aminosäuren im Bereich der aktiven Zentren lokalisiert sind. Ein aktives Zentrum ergibt sich dabei aus den drei konservierten Bereichen C1, C2 und C3 (vgl. Abbildung 3-29) verschiedener Untereinheiten (UE 1-3). Die in der Abbildung als UE1 bezeichnete Untereinheit enthält das konservierte Sequenzmotiv aus dem Bereich C3. Dieses ist auf einer flexiblen Schleife (*Loop*) lokalisiert. Das Threonin 161 aus Bereich C2 ist Bestandteil der Untereinheit UE2. Das Serin 114 ist auf der Untereinheit UE3 positioniert und befindet sich dem konservierten *Loop* gegenüber. Somit ist eine Beteiligung des Serins 114 am katalytischen Mechanismus nicht nur möglich sondern – in Anbetracht der nur wenigen konservierten Positionen – auch wahrscheinlich.

Zusammenfassend ist festzuhalten, daß die mittels Clusteranalyse gruppierten Enzyme ähnliche Reaktionen katalysieren, bei denen eine C-N- oder C-O-Bindung gespalten und – mit Ausnahme der Cycloisomerase – Fumarat als ein

Produkt gebildet wird. Obwohl der Grad der Sequenzähnlichkeit zwischen ihnen nur gering ist, gibt es drei konservierte Sequenzbereiche, die im Monomer weit auseinander liegen, im Tetramer jedoch eine Einheit bilden und die vier aktiven Zentren formen. Für einige der konservierten Aminosäuren konnte gezeigt werden, daß sie am Säure/Base-Mechanismus dieser Enzyme beteiligt sind.

3.7 Weitere Anwendungen der erhaltenen Klassifizierung

3.7.1 Erstellung neuer und Verbesserung bestehender Sequenzmotive

Die erhaltene Einteilung der Enzymsequenzen kann nicht nur dazu verwendet werden, um ähnliche enzymatische Reaktionsmechanismen zu identifizieren, sondern kann überdies auch zur Identifizierung invarianter Sequenzpositionen beitragen. Aus dem Vergleich von Aminosäuresequenzen verschiedener Proteine lassen sich wichtige Hinweise zur Aufklärung und Interpretation der Funktion gewinnen. Vor allem der Grad der Konservierung von Aminosäureresten innerhalb einer Familie oder Klasse kann zum Verstehen der Funktionsweise eines Proteins beitragen. Aufgrund ihrer Konservierung sind diese Aminosäuren meist von essentieller Bedeutung für die Funktion des Enzyms, da sie entweder die zur Katalyse erforderliche Struktur erhalten oder direkt an der Reaktion beteiligt sind. Anhand der generierten Sequenzcluster lassen sich multiple Alignments erstellen, mittels derer sich charakteristische Aminosäurepositionen und Sequenzmotive identifizieren lassen. Neben der Formulierung neuer Sequenzmotive können die Alignments auch dazu beitragen, bereits bestehende Motive zu präzisieren. Dies soll im folgenden am Beispiel der zuvor dargestellten Fumarat Lyasen illustriert werden (vgl. Abschnitt 3.6.3.5).

Die Qualität eines Sequenzmotivs bestimmt sich anhand zweier Kriterien:

- Die *Spezifität* ist ein Maß dafür, wie viele der mittels des Motivs gefundenen Sequenzen tatsächlich der entsprechenden Familie angehören (richtig Positive) und wie viele Sequenzen irrtümlich als Mitglieder dieser Familie identifiziert werden (falsch Positive). Sie ergibt sich aus dem Verhältnis: (richtig Positive / (richtig Positive + falsch Positive)).

- Die *Sensitivität* gibt an, welcher Anteil der zur Familie gehörenden Sequenzen mit Hilfe des Motivs identifiziert werden kann bzw. wie viele Familienmitglieder nicht durch das Motiv erfaßt werden. Sie errechnet sich aus dem Verhältnis: (richtig Positive / (richtig Positive + falsch Negative)).

Um eine möglichst gute Qualität zu erreichen, sollten mittels des Sequenzmotivs möglichst viele Familienmitglieder gefunden werden (hohe Sensitivität) und die Vorhersage möglichst fehlerfrei sein (hohe Spezifität).

Für die im Abschnitt 3.6.3.5 beschriebenen Fumarat Lyasen ist folgendes Motiv in PROSITE beschrieben (Kennummer: PS00163):

G-S-x(2)-M-x-{RS}-K-x-N

Diese Art der Darstellung liest sich wie folgt:
Zu Beginn des Motivs stehen ein Glycin und ein Serin. Darauf folgen zwei beliebige proteinogene Aminosäuren. Die nächste Position wird von einem Methionin eingenommen, woraufhin wieder eine beliebige Aminosäure folgt. Die geschweifte Klammer verweist darauf, daß an der nächsten Position alle außer den in Klammern genannten proteinogenen Aminosäuren folgen können. Darauf folgt ein Lysin, eine weitere beliebige Aminosäure und ein Asparagin.

Eine Suche dieses Sequenzmotivs gegen Swiss-Prot in Version 48.2 erbringt 280 Treffer in 280 verschiedenen Sequenzen. Von diesen gehören 274 der Familie der Fumarat Lyasen an. Sechs Treffer beziehen sich auf Sequenzen anderer Familien (falsch Positive). Des weiteren existieren 20 Mitglieder der Fumarat Lyase Familie, die nicht vom Motiv erfaßt werden, da sie in einer oder mehreren Positionen von diesem abweichen (falsch Negative). Somit beträgt die Spezifität dieses PROSITE-Motivs bei 97,86%, während die Sensitivität bei 93,20% liegt.

Anhand eines multiplen Alignments des eigenen, aus 295 Mitgliedern dieser Sequenzfamilie bestehenden Clusters, kann das in Tabelle 3-21 dargestellte Sequenzmotiv abstrahiert (vgl. auch Abschnitt 3.6.3.5).

Tabelle 3-21: Ausschnitt aus dem multiplen Alignment für die Familie der Fumarat Lyasen. Für jede EC-Klasse ist angegeben, welche Aminosäuren eine jeweilige Sequenzposition einnehmen können. „*X*“ steht für eine beliebige proteinogene Aminosäure.

EC-Klasse	Mögliche Aminosäuren einer jeweiligen Sequenzposition									
4.2.1.2	G	S	S	I	M	P	G	K	VIT	N
4.3.1.1	G	S	S	I	M	P	GA	K	VL	N
4.3.2.1	GIT	S	S	ILM	M	P	HQN	K	VKR	N
4.3.2.2	G	S	S	X	M	PA	HQY	K	X	N
5.5.1.2	GA	S	S	X	M	P	HQ	K	X	N
Konsensus	**GATI**	**S**	**S**	X	**M**	**PA**	X	**K**	X	**N**

Im PROSITE-Format dargestellt ergibt sich demnach folgendes Sequenzmotiv:

[GATI]-S-S-x-M-[PA]-x-K-x-N

Die eckigen Klammern geben an, daß die entsprechende Position ausschließlich von einer der in Klammern angegebenen Aminosäuren eingenommen werden kann.

Wird dieses Sequenzmotiv gegen Swiss-Prot in Version 48.2 verglichen, ergeben sich 286 Treffer in ebenso vielen Sequenzen. Alle mittels dieses Motivs gefunden Sequenzen gehören der betrachteten Familie der Fumarat Lyasen an. Es besteht somit keine Übereinstimmung zu Sequenzen anderer Familien. Lediglich acht bekannte Mitglieder der Familie konnten nicht mittels des Motivs detektiert werden. Die Spezifität des anhand der eigenen Klassifizierung erstellten Motivs beträgt somit 100%, während die Sensitivität bei 97,28% liegt.

Die Ursache für die vielfach bessere Qualität der erhaltenen Motive ist vor allem die größere Zahl der Sequenzen, die zu deren Erstellung verwendet wird. Viele der in PROSITE enthaltenen Motive wurden nur auf Basis relativ weniger Sequenzen erstellt. Daher gibt es mitunter eine Vielzahl von Familienmitgliedern, die dem postulierten Muster nicht entsprechen. Darüber hinaus werden bei der eigenen Methode nicht nur die Sequenzen eines Enzyms, sondern alle Mitglieder einer jeweiligen Sequenzfamilie verwendet. Dies erhöht nicht nur die Sequenzzahl, sondern erleichtert zudem die Identifizierung wichtiger Aminosäurepositionen, da bei homologen Sequenzen aus verschiedenen Enzym-

klassen häufig nur die für die Funktion essentiellen Residuen konserviert sind. Ein weiterer Vorteil der eigenen Methode besteht darin, daß ausschließlich die homologen Bereiche der Sequenzen zur Erstellung der multiplen Alignments verwendet werden. Verwendet man die vollständigen Sequenzen, so kann dies bei Multidomänenenzymen, die nur eine einzelne gemeinsame Domäne besitzen, in dürftigen und mitunter irreführenden Alignments resultieren. Alignments, die sich hingegen ausschließlich auf die ähnliche Domäne beschränken, können ein hohes Maß an Ähnlichkeit offenbaren.

3.7.2 Identifizierung und Korrektur falsch annotierter Sequenzen

Im Verlauf der Clusteranalyse wurden mehrere sequenzreiche Cluster erzeugt, die, abgesehen von einzelnen Ausnahmen, ausschließlich Enzyme einer einzigen EC-Klasse enthalten. Sucht man indes nach den übrigen Vertretern der davon abweichenden Sequenzen, so zeigt sich, daß diese jeweils in ein gemeinsames Cluster gruppiert worden sind. Es erscheint daher unwahrscheinlich, daß es sich bei den einzelnen Sequenzen, die einem anderen, ansonsten homogenen Cluster zugeordnet worden sind, um analoge Realisierungen der durch die jeweilige EC-Klasse beschriebenen Reaktion handelt. Es liegt vielmehr die Vermutung nahe, daß diese Sequenzen tatsächlich derselben EC-Klasse angehören wie die übrigen Sequenzen des jeweiligen Clusters, in den Datenbanken aber irrtümlich der falschen Enzymklasse zugeordnet worden sind. Eine Überprüfung der dazugehörigen Datenbankeinträge bestätigte dies. Auf diese Weise konnten insgesamt zwölf fehlerhaft annotierte Sequenzeinträge in Swiss-Prot und TrEMBL identifiziert werden. Bei den meisten Einträgen bezog sich der Irrtum lediglich auf die EC-Klassifikation, während alle übrigen Angaben korrekt waren. Bei einigen war jedoch die gesamte Zuordnung unzutreffend. Der Befund wurde mit entsprechenden Korrekturvorschlägen an das *Swiss Institute of Bioinformatics* übermittelt. Nach eingehender Prüfung seitens der dortigen Kuratoren wurden schließlich alle zwölf Vorschläge akzeptiert und entsprechend korrigiert (vgl. Tabelle 3-22).

Tabelle 3-22: Mit Hilfe der generierten Cluster wurden zwölf fehlerhaft annotierte Sequenzen identifiziert und entsprechende Korrekturvorschläge an das *Swiss Institute of Bioinformatics* übermittelt. Alle Änderungsvorschläge wurden von den dortigen Kuratoren akzeptiert und korrigiert.

Swiss-Prot- / TrEMBL-ID	vorherige EC-Klassifizierung	aktuelle EC-Klassifizierung
Q37183	1.1.1.1	4.1.1.39
Q8ELK1	1.1.1.38	5.1.1.1
Q7U5G1	2.3.3.13	2.2.1.6
O13326	2.4.1.49	2.5.1.49
P50125	2.4.1.49	2.5.1.49
Q7WTB4	3.1.2.23	3.2.1.23
Q8EL95	3.1.27.10	2.1.1.37
Q9FEW2	3.2.1.23	2.6.1.9
Q37186	4.1.1.29	4.1.1.39
Q7VH30	4.1.3.13	2.3.3.13
Q7VH30	4.1.3.132	2.3.3.13
Q8CX61	5.1.1.1	1.4.1.1

3.7.3 Funktionsvorhersage

Abschließend soll noch ein Blick auf die Möglichkeit geworfen werden, Vorhersagen über die mögliche Funktion neuer Proteine zu treffen. Durch großangelegte Sequenzierungsprojekte werden ständig große Mengen neuer Sequenzdaten ermittelt. Ihre enorme Zahl macht es unmöglich, die funktionelle Zuordnung der Sequenzen manuell durch Experten durchführen zu lassen. Für einen Großteil der erhaltenen Proteinsequenzen der bislang bekannten Genome ist die Funktion daher nicht bekannt. Einziger Ansatzpunkt für eine Analyse ist die Untersuchung der Proteinsequenz, welche mit den Sequenzen bereits charakterisierter Proteine verglichen wird, aus denen dann die Eigenschaften des neuen Proteins abgeleitet werden.

Im Gegensatz zu vergleichbaren Analysemethoden liefert die Clusteranalyse keine Zuordnungsvorschrift für künftige Beobachtungen. Um sie dennoch für prognostische Zwecke einsetzen zu können, muß sie für jede anstehende Vorhersage wiederholt werden. Dies ist aufgrund der dazu benötigten Rechenzeit nur für große Sequenzzahlen praktikabel. Es besteht aber dennoch die Möglichkeit, die generierten Daten auch zur Funktionsvorhersage einzelner Proteinsequenzen zu nutzen. Mit Hilfe der generierten Sequenzdomänensammlung können neue Sequenzen auf ihre Domänenzusammensetzung überprüft werden. Anhand des daraus resultierenden Domänenkataloges kann einer Sequenz dann eine potentielle Funktion zugeordnet werden.

Anhang

A. Verwendete Software und Datenbanken

1. Verwendete Programme:

Python 2.2	http://python.org/
MySQL 3.23.45	http://www.mysql.com/
BLAST 2.2.6	http://www.ncbi.nlm.nih.gov/Ftp/
Tcoffee 4.45	http://igs-server.cnrs-mrs.fr
MUSCLE 3.6	http://www.drive5.com/muscle/
Protein3Dfit	http://biotool.uni-koeln.de:8080/3dalign_neu/

2. Zusätzliche Module für Python:

mySQLdb 0.9	http://sourceforge.net/projects/mysql-python

3. Verwendete Datenbanken:

Swiss-Prot & TrEMBL	http://www.expasy.org/sprot/
IUBMB	http://www.chem.qmul.ac.uk/iubmb/enzyme/
BRENDA	http://www.brenda.uni-koeln.de/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
Prosite	http://www.expasy.org/prosite/
Protein Data Bank	http://www.rcsb.org/

B. Datenbankschema

Die aus öffentlichen Datenbanken abgerufenen Einträge und durch die verschiedenen Prozessierungsschritte entstandenen Daten wurden lokal in einer relationalen Datenbank abgelegt. Das Datenbankschema läßt sich grob in vier Bereiche gliedern, die in Abbildung B-1 entsprechend farblich hervorgehoben sind. Die Tabellenbezeichnungen sind kursiv angegeben:

- *Sequences* ist die zentrale Tabelle der Datenbank, über die alle übrigen Tabellen der Datenbank miteinander verknüpft sind. Sie enthält die Kerndaten zu jeder im Datensatz befindlichen Enzymsequenz wie zum Beispiel die Swiss-Prot-Kennummer (accession_number) und –ID (identification), anhand derer jede Sequenz eindeutig identifiziert werden kann, die Kurzbeschreibung des Enzyms (description), die dem deskriptiven Feld (DE) aus Swiss-Prot und TrEMBL entnommen wurde, sowie die Aminosäuresequenz (sequence).
- Die unter SWISS-PROT zusammengefaßten Tabellen enthalten darüber hinausgehende Sequenzinformationen, die aus Swiss-Prot und TrEMBL stammen und unter anderem die EC-Klassifikation (*ec_numbers)*, die Funktion (*functions*), aber auch andere Eigenschaften des Enzyms wie eventuell benötigt Cofaktoren (*cofactors*) betreffen. Swiss-Prot und TrEMBL enthalten überdies eine Vielzahl von Informationen aus externen Datenbanken, von denen einige (*pdb, pfam, prosite, intrepro)* ebefalls in die eigene Datenbank integriert wurden.
- Die Tabelle *alignments* enthält das Ergebnis der „alle-vs-alle"-Alignmentläufe. Zu jeder Sequenzpaarung ist der errechnete *E-value* (in der Form: -log(*E-value*) * 100), die Sequenz-ID der Suchsequenz (query_id) und des Datenbanktreffers (hit_id), die Länge beider Sequenzen (query_length, hit_length) und die miteinander alignierten Sequenzbereiche der beiden Sequenzen (query_start, query_end bzw. hit_start, hit_end) angegeben. Darüber hinaus ist angegeben, welche Grenzen für den jeweiligen Sequenzbereich während der Domänenvorhersage bestimmt wurden (qmod_start, qmod_end bzw. hmod_start, hmod_end).

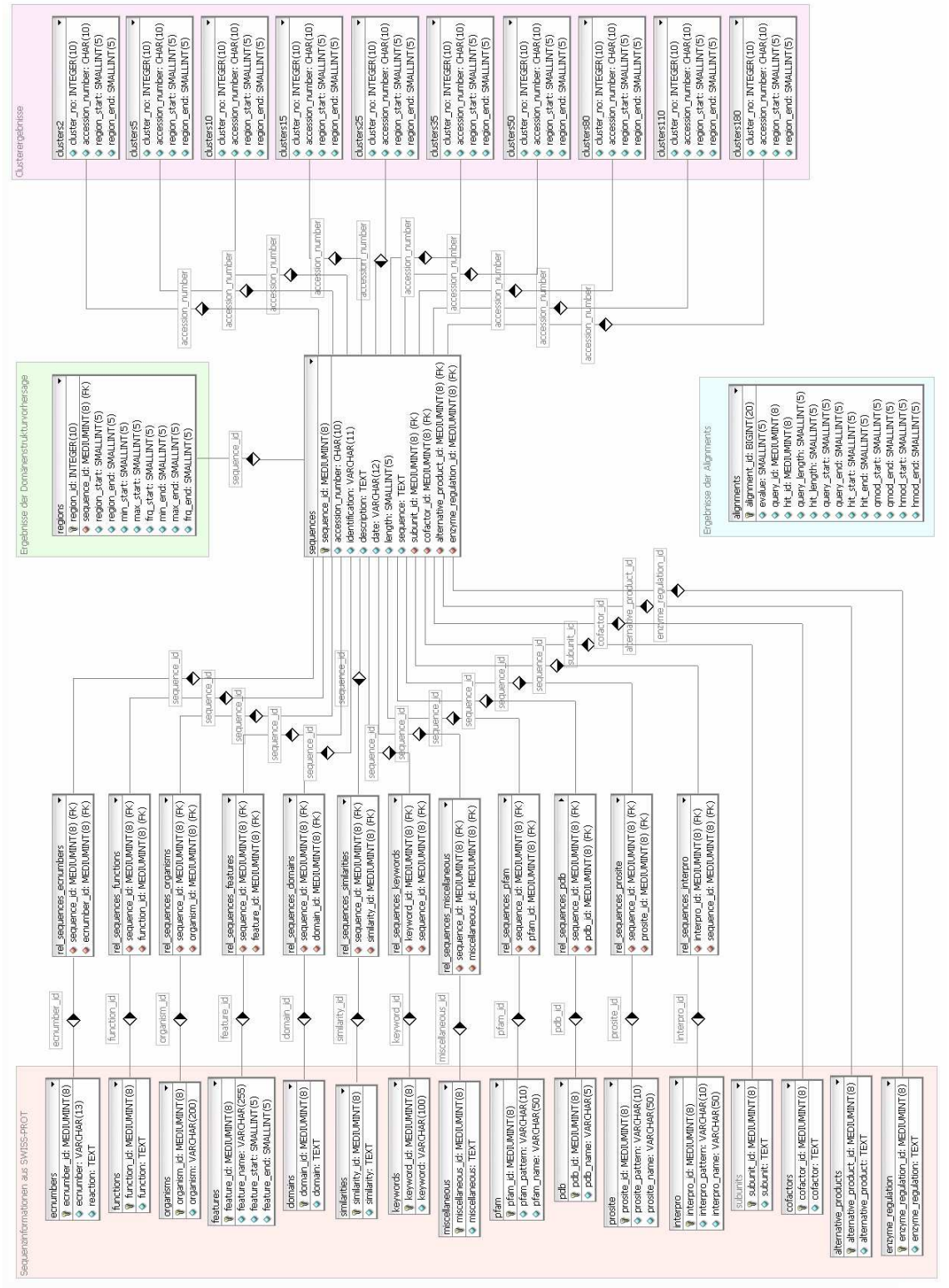


Abbildung B-1: Schema der Datenbank. Die Tabellen sind je nach Datenquelle oder Prozessierungsschritt farblich zusammengefaßt. Eine kurze Beschreibung der einzelnen Tabellen und Zusammenhänge findet sich im Text.

- Die Tabelle *regions* enthält alle nach der Domänenvorhersage zu einer Sequenz verbliebenen Alignmentbereiche. Diese bilden die Objektmenge der anschließenden Clusterung, bei der die Sequenzbereiche der verschiedenen Enzymsequenzen, die einander entsprechen, miteinander gruppiert werden.
- Die unter dem Begriff „Clusterergebnisse“ zusammengefaßten Tabellen enthalten schließlich das Resultat der Clusteranalyse. Es wurden Zwischenergebnisse für insgesamt zehn verschiedene Grenzwerte gespeichert (10^{-2}: *clusters2*, 10^{-5}: *clusters5*, etc.).

C. Inhalt der CD-ROM

Clusters/	Ergebnis der Clusteranalyse im txt-Format
Database/	Inhalt der aufgebauten Datenbank im MySQL-Format, Datenbankschema im MySQL-Format, Abbildung des Datenbankschemas im bmp-Format
IUBMB/	Datei zur Aktualisierung der Zuordnungen von Enzymklassen und Sequenzen
Posters/	Poster zur Dissertation
Reactions/	Datei zur Zuordnung von Enzymklasse und Reaktion
Source/	Quellen des entwickelten Programms
Statistic/	Dateien zur Häufigkeit der verschiedenen EC-Paarungen im txt-Format
Swiss-Prot & TrEMBL	Sequenzdatenbanken im txt-Format
Thesis/	Dissertation im pdf-Format
README	Informationen zu den enthaltenen Daten

D. Vorabveröffentlichungen

C. aus dem Spring, D. Schomburg (2002). *Sequence families of classified enzymes - Correlation with the function*. European Conference on Computational Biology, Saarbrücken. Posterpräsentation.

C. aus dem Spring, D. Schomburg (2003). *Prediction of protein domains and their boundaries from local sequence alignments*. German Conference on Bioinformatics, München. Posterpräsentation.

C. aus dem Spring, D. Schomburg (2003). *Mechanistic classification of enzymes in consideration of their modular architecture*. German Conference on Bioinformatics, München. Posterpräsentation.

C. aus dem Spring, D. Schomburg (2004*). Sequence-function-reaction relationships of classified enzymes*. BioPerspectives, Wiesbaden. Posterpräsentation.

C. aus dem Spring, D. Schomburg (2004). *Analysis of correlation between sequence homology and reaction mechanism in families of classified enzymes*. German Conference on Bioinformatics, Bielefeld. Posterpräsentation.

Literaturverzeichnis

[1] Voet D, Voet JG. 2002. Biochemie. Wiley-VCH.

[2] Tipton K, Boyce S. History of the enzyme nomenclature system. Bioinformatics. 2000;16(1):34-40.

[3] Nomenclature Committee of the International Union of Biochemistry. Recommendations of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes. Elsevier Publishing Company of Amsterdam 1964

[4] Hofmeister F. Über Bau und Gruppierung der Eiweißkörper. Erg Physiol. 1902;1:759-802.

[5] Anfinsen CR, Haber E, Sela E, White FH Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci. 1961;47:1309-1314.

[6] Pearson WR. Identifying distantly related protein sequences. Comput. Appl Biosci 1997;13(4):325-332.

[7] Yona G, Linial N, Tishby N, Linial M. A map of the protein space – an automatic hierarchical classification of all protein sequences. Ismb 1998;6:212-221.

[8] Patterson C. Homology in classical and molecular biology. Mol Biol Evol. 1988;5:603-625.

[9] Pearson WR. Effective protein sequence comparison. Methods Enzymol. 1996;266:227-258.

[10] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443-453.

[11] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;47(1):195-197.

[12] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA. 1988;85(8):2444-8.

[13] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

[14] Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol. 1991;219(3):555-65.

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

[16] Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9(1):56-68.

[17] Doolittle RF. Similar amino acid sequences: chance or common ancestry? Science. 1981;214(4517):149-59.

[18] Flores TP, Orengo CA, Moss DS, Thornton JM. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 1993;2(11):1811-26.

[19] Hilbert M, Bohm G, Jaenicke R. Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins. 1993;17(2):138-51.

[20] Doolittle RF. Reconstructing history with amino acid sequences. Protein Sci. 1992;1(2):191-200.

[21] Koonin EV, Tatusov RL, Rudd KE. Protein sequence comparison at genome scale. Methods Enzymol. 1996;266:295-322.

[22] Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol. 1997;273(1):349-54.

[23] Gerstein M. Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. Bioinformatics. 1998;14(8):707-14.

[24] Bolten E, Schliep A, Schneckener S, Schomburg D, Schrader R. Clustering protein sequences--structure prediction by transitive homology. Bioinformatics. 2001;17(10):935-41.

[25] Pipenbacher P, Schliep A, Schneckener S, Schonhuth A, Schomburg D, Schrader R. ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics. 2002;2:S182-91.

[26] Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol. 1996;6(3):377-85. Review.

[27] Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z,

[28] Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC. The Protein Information Resource. Nucleic Acids Res. 2003;31(1):345-7.

[29] Krause A, Vingron M. A set-theoretic approach to database searching and clustering. Bioinformatics. 1998;14(5):430-8.

[30] Krause A, Stoye J, Vingron M. Large scale hierarchical clustering of protein sequences. BMC Bioinformatics. 2005;6:15.

[31] Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science. 2003;300(5626):1701-3.

[32] Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature. 2002;420(6912):218-23.

[33] Doolittle RF, Bork P. Evolutionarily mobile modules in proteins. Sci Am. 1993;269(4):50-6.

[34] Gilbert W. Why genes in pieces? Nature. 1978;271(5645):501.

[35] Jensen, RA. Enzyme recruitment in evolution of new function. Annu Rev Microbiol. 1976;30:409-25.

[36] Babbitt PC, Gerlt JA. Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities. J Biol Chem. 1997;272(49):30591-4.

[37] O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform. 2002;3(3):275-84.

[38] Slater EC. IUB and IUBMB. IUBMB Life. 2005;57(4-5):379-80.

[39] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 2003;31(13):3784-8.

[40] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

[41] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89(22):10915-9.

[42] Altschul SF. A protein alignment scoring system sensitive at all evolutionary distances. J Mol Evol. 1993;36(3):290-300.

[43] Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem. 1994;18(3):269-85.

[44] Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 1996;266:554-71.

[45] Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. Nucleic Acids Res. 2006;1;34:D227-30.

[46] Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 1998;26(1):320-2.

[47] Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 2000;28(1):231-4.

[48] Attwood TK, Blythe MJ, Flower DR, Gaulton A, Mabey JE, Maudling N, McGregor L, Mitchell AL, Moulton G, Paine K, Scordis P. PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res. 2002;30(1):239-41.

[49] Pietrokovski S, Henikoff JG, Henikoff S. The Blocks database--a system for protein classification. Nucleic Acids Res. 1996;24(1):197-200.

[50] H iggins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673-4680.

[51] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302(1):205-17.

[52] Poirot O, O'Toole E, Notredame C. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Res. 2003;31(13):3503-6.

[53] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.

[54] Doolittle RF. On the trail of protein sequences. Bioinformatics. 2000;16(1):24-33.

[55] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986;5(4):823-6.

[56] Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12(2):85-94.

[57] Karp PD. What we do not know about sequence analysis and sequence databases. Bioinformatics. 1998;14(9):753-4.

[58] Shah I, Hunter L. Predicting enzyme function from sequence: a systematic appraisal. Proc Int Conf Intell Syst Mol Biol. 1997;5:276-83.

[59] Devos D, Valencia A. Practical limits of function prediction. Proteins. 2000;1;41(1):98-107.

[60] Rost B, Valencia A. Pitfalls of protein sequence analysis. Curr Opin Biotechnol. 1996 Aug;7(4):457-61.

[61] Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C. The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. J Mol Biol. 2001;24;311(4):693-708.

[62] Brenner SE. Errors in genome annotation. Trends Genet. 1999;15(4):132-3.

[63] Devos D, Valencia A. Intrinsic errors in genome annotation. Trends Genet. 2001;17(8):429-31.

[64] Russell RB, Sternberg MJ. Two new examples of protein structural similarities within the structure-function twilight zone. Protein Eng. 1997;10(4):333-8.

[65] Alexandrov NN, Fischer D. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. Proteins. 1996;25(3):354-65.

[66] Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998;26;95(11):6073-8.

[67] Musacchio A, Wilmanns M, Saraste M. Structure and function of the SH3 domain. Prog Biophys Mol Biol. 1994;61(3):283-97.

[68] Altschul SF, Boguski MS, Gish W, Wootton JC. Issues in searching molecular sequence databases. Nat Genet. 1994;6(2):119-29.

[69] National Center for Biotechnology Information. URL http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Seg.html

[70] Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res. 2001;1;29(1):55-7.

[71] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH - a hierarchic classification of protein domain structures. Structure. 1997;5(8):1093-108.

[72] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247(4):536-40.

[73] Orengo CA, Martin AM, Hutchinson G, Jones S, Jones DT, Michie AD, Swindells MB, Thornton JM. Classifying a protein in the CATH database of domain structures. Acta Crystallogr D Biol Crystallogr. 1998;54(Pt 6 Pt 1):1155-67.

[74] Corpet F, Gouzy J, Kahn D. The ProDom database of protein domain families. Nucleic Acids Res. 1998;26(1):323-6.

[75] Duncan K, Edwards RM, Coggins JR. The pentafunctional arom enzyme of Saccharomyces cerevisiae is a mosaic of monofunctional domains. Biochem J. 1987;246(2):375-86.

[76] Bae K, Mallick BK, Elsik CG. Prediction of protein interdomain linker regions by a hidden Markov model. Bioinformatics. 2005 May 15;21(10):2264-70.

[77] Sanger Institute. URL: http://www.sanger.ac.uk/Software/Pfam/help/region_help.shtml

[78] George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. Protein Eng. 2002;15(11):871-9.

[79] Knippers R. Molekulare Genetik. 1997. Thieme Georg Verlag

[80] Ghalambor MA, Heath EC. The metabolism of L-fucose: II. The enzymatic cleavage of L-fuculose 1-phosphate. J Biol Chem 1962;237, 2427-2433.

[81] Ghalambor MA, Heath EC. L-fuculose 1-phosphate aldolase. Methods Enzymol. 1966;9, 538-542.

[82] Sawada H, Takagi, Y. The metabolism of L-rhamnose in Eschericia coli: III. L-rhamnulosephosphate-aldolase. Biochim Biophys Acta. 1964;92, 26-32.

[83] Wolin MJ, Simpson FJ, Wood WA. L-Ribulose 5-phosphate D-xylulose 5-phosphate stereoisomerase and its role in L-arabinose fermentation. Biochim Biophys Acta. 1957;24, 635-638.

[84] Rutter WJ. Evolution of aldolases. Federation Proc. 1964;23,1248-1257.

[85] Joerger AC, Gosse C, Fessner WD, Schulz GE. Catalytic action of fuculose 1-phosphate aldolase (class II) as derived from structure-directed mutagenesis. Biochemistry 2000;39, 6033-6041.

[86] Kroemer M, Merkel I, Schulz GE. Structure and catalytic mechanism of L-rhamnulose-1-phosphate aldolase. Biochemistry. 2003;16;42(36):10560-8.

[87] Dreyer MK, Schulz GE. The spatial structure of the class II L-fuculose-1-phosphate aldolase from Escherichia coli. J Mol Biol. 1993;231, 549-553.

[88] Lessel U, Schomburg D. Similarities between protein 3-D structures. Protein Eng. 1994;7(10):1175-87.

[89] Dreyer MK, Schulz GE. Catalytic mechanism of the metal-dependent fuculose aldolase from Escherichia coli as derived from the structure. J Mol Biol. 1996;259, 458-466.

[90] Dreyer MK, Schulz GE. Refined high-resolution structure of the metal-ion dependent L-fuculose-1-phosphate aldolase (class II) from Escherichia coli. Acta Crystallogr Sect. 1996;D 52, 1082-1091.

[91] Lee N, Gielow W, Martin R, Hamilton E, Fowler A. The organization of the araBAD operon of Escherichia coli. Gene 1986;47, 231-244.

[92] Kopp J, Kopriva S, Süss KH, Schulz GE. Structure and mechanism of the amphibolic enzyme D-ribulose-5-phosphate 3-epimerase from potato chloroplasts. J Mol Biol 1999;287, 761-771.

[93] Deupree JD, Wood WA. L-Ribulose 5-phosphate 4-epimerase from Aerobacter aerogenes. Evidence for nicotinamide adenine dinucleotide-independent 4-epimerization by the crystalline enzyme. J Biol Chem. 1970;245, 3988-3995.

[94] Thoden JB, Frey PA, Holden HM. Molecular structure of the NADH/UDP-glucose abortive complex of UDP-galactose 4-epimerase from Escherichia coli: implications for the catalytic mechanism. Biochemistry 1996;35, 5137-5144.

[95] Deupree JD, Wood WA. L-Ribulose 5-phosphate 4-epimerase from Aerobacter aerogenes. Evidence for a role of divalent metal ions in the epimerization reaction. J Biol Chem. 1972;247, 3093-3097.

[96] Davis L, Lee N, Glaser L. On the mechanism of the pentose phosphate epimerases. J Biol Chem 1972;247, 5862-5866.

[97] Johnson AE, Tanner ME. Epimerization via carbon-carbon bond cleavage. L-Ribulose-5-phosphate 4-epimerase as a masked class II aldolase. Biochemistry 1998;37, 5746-5754.

[98] Luo Y, Samuel J, Mosimann SC, Lee JE, Tanner ME, Strynadka NCJ. The structure of L-ribulose-5-phosphate-4-epimerase: an aldolase-like platform for epimerization. Biochemistry 2001;40,14763-14771.

[99] Lee LV, Poyner RR, Vu MV, Cleland WW. Role of metal ions in the reaction catalyzed by L-ribulose-5-phosphate 4-epimerase. Biochemistry 2000;39, 4821-4830.

[100] Samuel J, Luo Y, Morgan PM, Strynadka NCJ, Tanner ME. Catalysis and binding in L-ribulose-5-phosphate-4-epimerase: a comparison with L-fuculose-1-phosphate-aldolase. Biochemistry 2001;40,14772-14780.

[101] Tanner ME. Understanding nature's strategies for enzyme-catalyzed racemization and epimerization. Acc Chem. 2002;35, 237-246.

[102] Samuel J, Tanner ME. Mechanistic aspects of enzymatic carbohydrate epimerization. Nat Prod. 2002;Rep. 19, 261-277.

[103] Dincturk HB, Knaff DB. The evolution of glutamate synthase. Mol Biol Rep. 2000;27(3):141-8.

[104] Temple SJ, Vance CP, Gnatt JS. Glutamate synthase and nitrogen assimilation. Trends Plant Sci. 1998;3:51-56.

[105] Vollmer SJ, Switzer RL, Hermodson MA, Bower SG, Zalkin H. The glutamine-utilizing site of Bacillus subtilis glutamine phosphoribosylpyrophosphate amidotransferase. J Biol Chem. 1983;258(17):10582-5.

[106] Cooper AJ, Meister A. Isolation and properties of a new glutamine transaminase from rat kidney. J Biol Chem. 1974;249(8):2554-61.

[107] Calcagno M, Levy JA, Arrambide E, Mizraji E. L-glutamine D-fructose-6-phosphate amidotransferase of chick cartilage. Evidence for a random mechanism. Enzymologia. 1971;41(3):174-82.

[108] Richards NG, Schuster SM. An alternative mechanism for the nitrogen transfer reaction in asparagine synthetase. FEBS Lett. 1992;313(2):98-102.

[109] Zalkin H, Smith JL. Enzymes utilizing glutamine as an amide donor. Adv Enzymol Relat Areas Mol Biol. 1998;72:87-144.

[110] Massiere F, Badet-Denisot MA. The mechanism of glutamine-dependent amidotransferases. Cell Mol Life Sci. 1998;54(3):205-22.

[111] Buchanan JM. The amidotransferases. Adv Enzymol Relat Areas Mol Biol. 1973;39:91-183.

[112] Zalkin H. The amidotransferases. Adv Enzymol Relat Areas Mol Biol. 1993;66:203-309.

[113] Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, et al. The alpha/beta hydrolase fold. Protein Eng. 1992;5(3):197-211.

[114] Brannigan JA, Dodson G, Duggleby HJ, Moody PC, Smith JL, Tomchick DR, Murzin AG. A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. Nature. 1995;378(6555):416-9.

[115] van den Heuvel RH, Ferrari D, Bossi RT, Ravasio S, Curti B, Vanoni MA, Florencio FJ, Mattevi A. Structural studies on the synchronization of catalytic centers in glutamate synthase. J Biol Chem. 2002;277(27):24579-83.

[116] Smith JL. Structures of glutamine amidotransferases from the purine biosynthetic pathway. Biochem Soc Trans. 1995;23(4):894-8.

[117] Thoden JB, Miran SG, Phillips JC, Howard AJ, Raushel FM, Holden HM. Carbamoyl phosphate synthetase: caught in the act of glutamine hydrolysis. Biochemistry. 1998;37(25):8825-31.

[118] van den Heuvel RH, Svergun DI, Petoukhov MV, Coda A, Curti B, Ravasio S, Vanoni MA, Mattevi A. The active conformation of glutamate synthase and its binding to ferredoxin. J Mol Biol. 2003;330(1):113-28.

[119] Saraste M, Sibbald PR, Wittinghofer A. The P-loop--a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci. 1990;15(11):430-4.

[120] Schaap D, van der Wal J, van Blitterswijk WJ. Consensus sequences for ATP-binding sites in protein kinases do not apply to diacylglycerol kinases. Biochem J. 1994;304 (Pt 2):661-2.

[121] Conti E, Franks NP, Brick P. Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. Structure. 1996;4(3):287-98.

[122] Coutinho PM, Henrissat B. Carbohydrate-active enzymes: an integrated database approach. In "Recent Advances in Carbohydrate Bioengineering", H.J. Gilbert, G. Davies, B. Henrissat and B. Svensson eds., The Royal Society of Chemistry, Cambridge, 1999;pp. 3-12.

[123] MacGregor EA, Janecek S, Svensson B. Relationship of sequence and structure to specificity in the alpha-amylase family of enzymes. Biochim Biophys Acta. 2001;1546(1):1-20.

[124] Henrissat B, Davies GJ (1997) Structural and sequence-based classification of glycoside hydrolases. Curr. Op. Struct. Biol. 7:637-644

[125] Sinnott ML. Catalytic mechanisms of enzymic glycoside transfer. Chem Rev. 1990;90:1170-1202.

[126] McCarter JD, Withers SG. Mechanisms of enzymatic glycoside hydrolysis. Curr. Opin. Struct. Biol 1994;4:885-892

[127] Woods, S. A., J. S. Miller und J. R. Guest. Sequence homologies between argininosuccinase, aspartase, and fumarase: A family of structurally-related enzymes. FEMS Microbiol. Lett. 1988;51:181-186.

[128] Woods SA, Schwartzbach SD, Guest JR. Two biochemically distinct classes of fumarase in Escherichia coli. Biochim Biophys Acta. 1988;954(1):14-26.

[129] Falzone CJ, Karsten WE, Conley JD, Viola RE. L-aspartase from Escherichia coli: substrate specificity and role of divalent metal ions. Biochemistry. 1988;27(26):9089-93.

[130] O´Brien WE, Barr RH. Argininosuccinate lyase: purification and characterization from human liver. Biochemistry. 1981;20:2056-2060.

[131] Stone RL, Zalkin H, Dixon JE. Expression, purification, and kinetic characterization of recombinant human adenylosuccinate lyase. J Biol Chem. 1993;268(26):19710-6.

[132] Williams SE, Woolridge EM, Ransom SC, Landro JA, Babbitt PC, Kozarich JW. 3-Carboxy-cis,cis-muconate lactonizing enzyme from Pseudomonas putida is homologous to the class II fumarase family: a new reaction in the evolution of a mechanistic motif. Biochemistry. 1992;31(40):9768-76.

[133] Hansen JN, Dinovo EC, Boyer PD. Initial and equilibrium ^{18}O, ^{14}C, ^{3}H, and ^{2}H exchange rates as probes of the fumarase reaction mechanism. J Biol Chem. 1969 Nov 25;244(22):6270-9.

[134] Porter DJT, Bright HJ. 3-Carbanionic substrate analogoues bind very tightly to fumarase and aspartase. J Biol Chem. 1980;255:4772-4780.

[135] Rose IA. How fumarase recycles after the malat --> fumarate reaction. Insight into the reaction mechanism. Biochemistry. 1998;37(51):17651-8.

[136] Alberty RA. Fumarase. The Enzymes, 2nd Ed (Boyer PD, Lardy H, Myrbäck K, eds.) 1961;5: 531-544.

[137] Colowick SP, Englard S. On the mechanism of an anaerobic exchange reaction catalyzed by succinic dehydrogenase preparations. J Biol Chem. 1956;221(2):1019-35.

[138] Englard S, Colowick SP. On the mechanism of an anaerobic exchange reaction catalyzed by succinic dehydrogenase preparations. Science. 1955;121(3155):866-7.

[139] Hill RL, Teipel JW. Fumarase and crotonase. The Enzymes, 3rd Ed. (Boyer PD, ed.) 1971;5,539-571.

[140] Massey V. Studies on fumarase. 4. The effects of inhibitors on fumarase activity. Biochem J. 1953;55(1):172-7.

[141] Alberty RA, Miller WG, Fisher HF. J. Am. Chem. Soc. 1957;79:3973.

[142] Bradshaw RA, Robinson GW, Hass GM, Hill RL. The reaction of fumarase with iodoacetate and 4-bromocrotonate. J Biol Chem. 1969;244(7):1755-63.

[143] Piatigorsky J, O'Brien WE, Norman BL, Kalumuck K, Wistow GJ, Borras T, Nickerson JM, Wawrousek EF. Gene sharing by delta-crystallin and argininosuccinate lyase. Proc Natl Acad Sci U S A. 1988;85(10):3479-83.

[144] Mori M, Matsubasa T, Amaya Y, Takiguchi M. Molecular evolution from argininosuccinate lyase to delta-crystallin. Prog Clin Biol Res. 1990;344:683-99.

[145] Barbosa P, Wistow GJ, Cialkowski M, Piatigorsky J, O'Brien WE. Expression of duck lens delta-crystallin cDNAs in yeast and bacterial hosts.Delta 2-crystallin is an active argininosuccinate lyase. J Biol Chem. 1991 Nov 25;266(33):22319-22.

[146] Lee HJ, Chiou SH, Chang GG. Biochemical characterization and kinetic analysis of duck delta-crystallin with endogenous argininosuccinate lyase activity. Biochem J. 1992;283 (Pt 2):597-603.

[147] Kondoh H, Araki I, Yasuda K, Matsubasa T, Mori M. Expression of the chicken 'delta 2-crystallin' gene in mouse cells: evidence for encoding of argininosuccinate lyase. Gene. 1991;99(2):267-71.

[148] Chiou SH, Hung CC, Lin CW. Biochemical characterization of crystallins from pigeon lenses: structural and sequence analysis of pigeon delta-crystallin. Biochim Biophys Acta. 1992;1160(3):317-24.

[149] Simpson A, Bateman O, Driessen H, Lindley P, Moss D, Mylvaganam S, Narebor E, Slingsby C. The structure of avian eye lens delta-crystallin reveals a new fold for a superfamily of oligomeric enzymes. Nat Struct Biol. 1994;1(10):724-34.

[150] Weaver TM, Levitt DG, Donnelly MI, Stevens PP, Banaszak LJ. The multisubunit active site of fumarase C from Escherichia coli. Nat Struct Biol. 1995;2(8):654-62.

[151] Abu-Abed M, Turner MA, Vallee F, Simpson A, Slingsby C, Howell PL. Structural comparison of the enzymatically active and inactive forms of delta crystallin and the role of histidine 91. Biochemistry. 1997;36(46):14012-22.

[152] Patejunas G, Barbosa P, Lacombe M, O'Brien WE. Exploring the role of histidines in the catalytic activity of duck delta-crystallins using site-directed mutagenesis. Exp Eye Res. 1995;61(2):151-4.

[153] Turner MA, Simpson A, McInnes RR, Howell PL. Human argininosuccinate lyase: a structural basis for intragenic complementation. Proc Natl Acad Sci U S A. 1997;94(17):9063-8.

[154] Lee TT, Worby C, Bao ZQ, Dixon JE, Colman RF. His68 and His141 are critical contributors to the intersubunit catalytic site of adenylosuccinate lyase of Bacillus subtilis. Biochemistry. 1999 Jan 5;38(1):22-32.

[155] Sampaleanu, L. M. F. Vallée, G. D. Thompson und P. L. Howell. Three-dimensional structure of the argininosuccinate lyase frequently complementing allele Q286R. Biochemistry. 2001;40:15570-15580.

[156] Sampaleanu LM, Yu B, Howell PL. Mutational analysis of duck delta 2 crystallin and the structure of an inactive mutant with bound substrate provide insight into the enzymatic mechanism of argininosuccinate lyase. J Biol Chem. 2002;277(6):4166-75.

[157] Sampaleanu LM, Codding PW, Lobsanov YD, Tsai M, Smith GD, Horvatin C, Howell PL. Structural studies of duck delta2 crystallin mutants provide insight into the role of Thr161 and the 280s loop in catalysis. Biochem J. 2004 Dec 1;384(Pt 2):437-47.

[158] Brosius JL, Colman RF. Three subunits contribute amino acids to the active site of tetrameric adenylosuccinate lyase: Lys268 and Glu275 are required. Biochemistry. 2002 Feb 19;41(7):2217-26.

[159] Shi W, Dunbar J, Jayasekera MM, Viola RE, Farber GK. The structure of L-aspartate ammonia-lyase from Escherichia coli. Biochemistry. 1997 Jul 29;36(30):9136-44.