DIE *LIPASE ENGINEERING DATABASE* SYSTEMATISCHE ANALYSE FAMILIENSPEZIFISCHER EIGENSCHAFTEN UND DER SEQUENZ-STRUKTUR-FUNKTIONSBEZIEHUNG VON α/β-HYDROLASEN

VON DER FAKULTÄT CHEMIE DER UNIVERSITÄT STUTTGART
ZUR ERLANGUNG DER WÜRDE EINES
DOKTORS DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
GENEHMIGTE ABHANDLUNG

VORGELEGT VON

MARKUS FISCHER
AUS LUDWIGSBURG (Württ.)

HAUPTBERICHTER: PROF. DR. ROLF D. SCHMID
MITBERICHTER: PROF. DR. DIETER H. WOLF
VORSITZENDER: PROF. DR. EMIL RODUNER
TAG DER MÜNDLICHEN PRÜFUNG: 15. FEBRUAR 2005

INSTITUT FÜR TECHNISCHE BIOCHEMIE DER UNIVERSITÄT STUTTGART

2004

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Hilfsmittel und Literatur angefertigt habe.

Stuttgart, den 20.11.2004 Markus Fischer

DIE NATUR IST SO GEMACHT, DASS SIE VERSTANDEN WERDEN KANN. ODER VIELLEICHT SOLLTE ICH RICHTIGER SAGEN, UNSER DENKEN IST SO GEMACHT, DASS ES DIE NATUR VERSTEHEN KANN.

(Werner Heisenberg)

Danksagung

Prof. Rolf D. Schmid danke ich für die Überlassung des interessanten Themas und die hervorragenden Arbeitsbedingungen am Institut für Technische Biochemie. In den vergangenen Jahren hatte ich die Gelegenheit viel von ihm zu lernen.

Mein besonderer Dank gilt Herrn Priv. Doz. Dr. Jürgen Pleiss für die ausgezeichnete Betreuung dieser Arbeit. Dank seiner Anleitung wurde mein Interesse an der akademischen Forschung geweckt, der ich auch in Zukunft verhaftet bleiben möchte.

Wesentliche Hilfe für mich waren die konstruktive Kritik und Diskussionsbereitschaft von Herrn Dr. Holger Scheib und dafür gilt mein herzlichster Dank.

Für die anregenden Diskussionen nicht nur naturwissenschaftlicher Themen in gemütlicher Runde danke ich den langjährigen Freunden Herrn Dr. Ulrich Kahlow, Herrn Dipl. Chem. Christoph Kaiser, Herrn Dipl. Inf. Robert Dress und Frau Dipl. Phys. Anja Mindermann.

Bei Frau Dr. Sandra Barth möchte ich mich für die wissenschaftliche Zusammenarbeit bedanken. Gemeinsam bestritten wir die Betreuung vieler Praktika und so manche unterhaltsame Nachsitzung.

Herrn Florian Barth, Frau Dipl. Lebensmittelchem. Nicole Beuttenmüller und Frau Dipl. Biol. Monika Rusnak danke ich für die vielen kulinarischen Gaumenfreuden, die wir gemeinsam genießen durften.

Danken möchte ich auch Herrn Dr. Erik Henke und Frau Dr. Sandra Vorlová, für eine über Jahre und Ländergrenzen hinweg gepflegte Freundschaft.

Den Mitgliedern der Bioinformatikgruppe möchte ich für das mehr als angenehme Arbeitsklima danken. Stellvertretend für die, die mich im Laufe meines Werdegangs in dieser Arbeitsgruppe begleitet haben, möchte ich insbesondere den jungen Wilden Dipl. Biol. (t.o.) Fabian Bös, Dipl. Biol. (t.o.) Michael Knoll, Dipl. Chem. Stephan Tatzel, Dipl. Chem. Peter Trodler und Dipl. Biol. (t.o.) Alexander Steudle danken.

Und gedankt sei auch all den treuen Freunden, die stets daran gezweifelt haben, dass dieses Werk jemals seine Vollendung findet.

Nicht zuletzt danke ich meinen Eltern und Geschwistern für die geborgene Atmosphäre, die stets eine Zuflucht vor dem hektischen Alltag war und so nicht minder zum Gelingen dieser Arbeit beitrug.

In Erinnerung An

Dr. Ulrich Kahlow

Inhalt

Abkürzungsverzeichnis

a	Jahr(e)
AchE	Acetylcholinesterase
APS	Ammoniumpersulfat
Amp	Ampicillin
ASE	*Alcaligenes sp.* Esterase
BCA	Bicinchoninsäure
BCL	*Burkholderia cepacia* Lipase
BGL	*Burkholderia glumae* Lipase
bp	Basenpaare
BSA	Rinderserumalbumin
BSD	*Bacillus subtilis* Deacetylase
B. subtilis	*Bacillus subtilis*
c	Konzentration
C	Celsius
CALB	*Candida antarctica* Lipase B
C.elegans	*Caenorhabditis elegans*
CRL	*Candida rugosa* Lipase
CVL	*Chromobacterium viscosum* Lipase
D	Tag(e)
DBMS	Database Management System
DGL	*Canis familiaris* Lipase
dH2O	Destilliertes Wasser
ddH2O	Doppelt destilliertes Wasser
DMSO	Dimethylsulfoxid
dNTP	2´-Desoxynucleosid-5´-triphosphat
DNA	Desoxyribonucleinsäure
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
dsDNA	doppelsträngige DANN
DWARF	Data Warehouse for Analysing Protein Families
E.C.	Enzyme Commission
E. coli	*Escherichia coli*
EDTA	Ethylendiamintetraacetat
FDBS	Federated Database System
FSC	*Fusarium solani* Cutinase
Glc	Glucose
GPL	*Cavia porcellus* Lipase
h	Stunde(n)
HDBMS	Hierarchical Database Management System
HLL	*Humicola lanuginosa* Lipase
HMM	Hidden Markov Model
HPL	*Equus caballus* Lipase
HuPL	Humane Pankreaslipase
IPTG	Isopropyl-β-D-thio-galactopyranosid
kB	Kilobase (1000 Basenpaare)
kcat	Wechselzahl
kDa	Kilo-Dalton
KM	Michaelis-Konstante
KOAc	Kaliumacetat

KP_i	Kaliumphosphat-Puffer mit unterschiedlichem Verhältnis von Kaliumhydrogenphosphat und Kaliumdihydrogenphosphat
LB	Luria-Bertani (Komplexmedium)
LB-Amp	Ampicillinhaltiges LB-Medium
LED	Lipase Engineering Database
min	Minute(n)
M	Mol pro Liter
ML	Maximum Likelihood
Mr	Molekulargewicht
mM	Millimol pro Liter
mRNA	messenger-RNA
MSHFBA	N-Methyl-N-trimethylsilyl-heptafluorobutyramid
(m/v)	Masse / Volumen
m.w.N.	mit weiteren Nachweisen
NaOAc	Natriumacetat
NJ	Neighbour Joining
Nu	katalytisch aktives Nucleophil
OD_{xxx}	Optische Dichte bei einer Wellenlänge von xxx nm
OODBMS	Object orientated Database Management System
PAGE	Polyacrylamid-Gelelektrophorese
PEG	Polyethylenglykol
PCB	Polychlorierte Biphenyle
PCR	Polymerasekettenreaktion
PeCL	*Penicillium camemberti* Lipase
Pfu	*Pyrococcus furiosus*
PPC	*Pseudomonas putida* Carboxymethylenbutenolidase
PSSM	Position Specific Scoring Matrices
RDBMS	Relational Database Management System
RDL	*Rhizopus delemar* Lipase
R.delemar	*Rhizopus delemar*
RML	*Rhizomucor miehei* Lipase
RMSD	Root Mean Square Deviation
RNA	Ribonucleinsäure
ROL	*Rhizopus oryzae* Lipase
R.oryzae	*Rhizopus oryzae*
Rt	Retentionszeit
RT	Raumtemperatur
SAB	*Streptomyces aureofaciens* Bromoperoxidase
S. cerevisiae	*Saccharomyces cerevisiae*
SDS	Natriumdodecylsulfat
SEL	*Streptomyces exfoliatus* Lipase
SKP	Streng konservierte Positionen
sp.	Spezies
TAE	Tris-Acetat-EDTA
Taq	*Thermus aquaticus*
TBE	Tris-Borsäure-EDTA
TE	Tris-EDTA
TEMED	N,N,N´,N´-Tetramethylethylendiamin
TfB	Transformation Buffer
THF	Tetrahydrofuran
Tris	Tris-(hydroxymethyl)-aminomethan

TSS	Transformation storage solution
U	Unit (Stoffmengenumsatz in µmol pro Minute)
ÜN	Über Nacht
UPGMA	Unweighted pair group method using arithmetic mean
Upm	Umdrehungen pro Minute
UV	Ultraviolett
VIS	sichtbarer Bereich des Lichts
(v/v)	Volumen / Volumen
(w/v)	Gewicht / Volumen
WT	Wildtyp
µM	Mikromol pro Liter
ZKP	Zwingend konservierte Positionen

Kurzzusammenfassung

Im Rahmen dieser Arbeit wurde die Familie der α/β-Hydrolasen systematisch untersucht. Grundlage dieser Analyse war die Entwicklung eines Data Warehouse Systems zur Integration von Proteinsequenz- und Strukturdaten in Verbindung mit funktionell relevanten Informationen für große Proteinfamilien. Als Anwendung dieses Data Warehouse Systems wurde die *Lipase Engineering Database* (LED) erstellt, die alle Mitglieder der α/β-Hydrolasen enthält. Das Release 2.3 der LED enthält 3148 Sequenzeinträge für 2313 Proteine, wobei 35% der Proteine als putativ definiert sind. Für 96 Proteineinträge sind 261 Struktureinträge in der LED abgelegt. Aufgrund der Sequenzähnlichkeit wurden die α/β-Hydrolasen in 37 Superfamilien und 103 homologe Familien eingeteilt.

Die LED diente zur systematischen Identifikation familienspezifischer Deskriptoren auf Sequenz- und Strukturebene. Die Bestimmung streng konservierter Positionen für acht repräsentative α/β-Hydrolasesuperfamilien half bei der Identifikation des Faltungsnukleus, der den α/β-Hydrolase Fold charakterisiert. Daraus konnte ein Modell für die Faltung dieser Proteine skizziert werden.

Anhand der Architektur des funktionell wichtigen *oxyanion holes* konnten diese in drei Klassen eingeteilt werden: (1) die GGGX Klasse, für die die *oxyanion hole* bildende Aminosäure am C-terminalen Ende des hoch konservierten Musters GGG lokalisiert ist und von einer streng konservierten, hydrophoben Aminosäure X gefolgt wird, (2) die GX Klasse, für die die *oxyanion hole* Aminosäure X einem streng konserviertem Glycin folgt und (3) die Y Klasse deren *oxyanion hole* durch die Hydroxylgruppe eines streng konservierten Tyrosin gebildet wird. Für die strukturelle Stabilisierung des *oxyanion holes* der GX Klasse wurde die Wechselwirkung der Seitenkette X mit Anker Aminosäuren identifiziert. Dieses Anker-Konzept konnte experimentell durch den Austausch des Anker-Moduls in der Lipase aus *Burkholderia cepacia* gegen das der Lipase aus *Rhizomucor miehei* bestätigt werden.

Die Form und die physikalisch-chemischen Eigenschaften der Fettsäurebindungsstelle von acht α/β-Hydrolasen wurden untersucht, um deren Substratspezifität auf molekularer Ebene nachzuvollziehen. Diese konnten so in vier Gruppen eingeteilt werden und half bei der Identifikation von Aminosäuren, die für die Vermittlung der Substratkettenlängenspezifität verantwortlich sind.

Für die Superfamilie abH1 wurden mit der in CODEHOP implementierten Methode degenerierte familienspezifischen Primer erstellt, die sich zur Identifikation neuer Carboxylesterasen in aus Bodenproben isolierter DNA eignen.

Abstract

The emergence of the 'omics' era has redefined the field of modern biology dramatically. The methods of the functional and structural genomics lead to a rapidly increasing amount of sequence and structure data. The field of Bioinformatics provide the methods to utilize these data for the *a priori* determination of proteins' function assuming that similar proteins are evolutionary related,[1] thus sharing the same protein fold,[2] resulting in a functional relationship.

However this deductive approach for predicting the function of proteins requires a tight integration of vast amounts of heterogenous data. To facilitate the systematic analysis of large protein families in this study a data warehouse system to store and analyze sequence, structure, and functional annotation information was developed. The data model is organized in three sections: protein, protein sequence, and protein structure. Based on multisequence alignments and phylogenetic analyses, homologous families and superfamilies are defined. Protein sequences are aligned, protein structures superposed, and functionally relevant residues are consistently annotated.

This data warehouse system was applied to set up the Lipase Engineering Database (LED) which includes all members of the α/β hydrolase fold family. The α/β hydrolases represent one of the largest families in protein space of structurally related proteins. They are a member of the doubly wound superfold[3,4] and contain a central mostly parallel β-sheet consisting of eight β-strands.[5] To both sides of this β-sheet amphiphilic α-helices are packed closely. The α/β hydrolases catalyze a broad variety of chemical reactions and accept highly different substrates,[6] but all share the same catalytic machinery, which consist of a nucleophilic-His-acid triad and the oxyanion hole stabilizing the tetrahedral substrate intermediate. Twelve years ago, structure data for only five homologous proteins were known.[5] In the meantime the protein structure classification database CATH[7] lists 333 different structure data sets classified in 44 sequence families. Many members of this family have been investigated for their catalytic and non-catalytic properties intensively.

In the release 2.3 the LED includes 3148 sequences of 2313 proteins entries, 35% of them being putative proteins. For 96 protein entries 261 structure data sets from the Protein Databank were stored. Proteins were assigned to 37 superfamilies and 103 homologous families. The sequence annotation validation resulted in the assignment of the catalytic triad for all superfamilies, for which structure information was available. By sequence comparison,

these functional residues were assigned in 25% of homologous families and 53% of all sequence entries.

The LED has been applied to systematically analyze the α/β hydrolase fold family. Based on sequence and structure data family specific descriptors were inferred, the effect of mutations could be explained, new functionally relevant modules were identified and substrate specificity could be predicted by the gene sequence.

Using the sum-of-pairs method (SP) for calculating the conservation of residues within multisequence alignments the most conserved amino acids for eight superfamilies were detected. Since the sequence similarity between these superfamilies is not sufficient for a direct sequence comparison based on a multisequence alignment, a structure alignment was generated to map position specific information between these superfamilies. By this approach the mutually conserved residues were identified. These residues are conserved within families, but show differences between families though they are located at equal positions within a common protein fold. These positions are claimed to form the folding nucleus. For the α/β hydrolase fold 54 mutually conserved residues were identified. For these position the hydrophobic residues were significantly overrepresented compared to the background frequencies of amino acids within the LED. 83% of these residues were not accessible for the solvent and were buried at the hydrophobic core of the α/β hydrolases. They mainly occupy so called supersecondary structures (SSS). These SSS represent energetically preferred elements which initialize the local folding within the protein.[8] Based on this findings a model for the process of folding of the α/β hydrolases was postulated.

To estimate the relationships between the superfamilies of α/β hydrolases, Hidden Markov Models (HMM) were generated for each homologous family. Distances were inferred by profile HMM-HMM comparison based on p-values.[9] These distances were used to create a phylogenetic tree based on the UPGMA method as implemented in the PHYLIP package.[10] According to this tree a systematic nomenclature for α/β hydrolases was introduced: abHn.m, where abH is followed by the superfamily number n, which is separated from the homologous family number m by a dot.

Based on structure and sequence analysis of the oxyanion hole, α/β hydrolases were classified into three classes, the GGGX-, GX- and Y-class.

The GX-class consists of 20 superfamilies with known protein structures, where the oxyanion hole forming residue X is structurally well conserved and was preceded by a strictly conserved glycine. Based on profile HMM-HMM comparisons another seven superfamilies, which contain no protein structure data, were also assigned to the GX class and show

significant similarity in the oxyanion hole region. This class comprises the families abH7-abH26, abH31-abH37.

The GGGX-class consists of five superfamilies with known protein structures, where the oxyanion hole forming residue is located in a well conserved GGG pattern, which is followed by a conserved hydrophobic amino acid X. The backbone amide of glycine G preceding the residue X is forming the oxyanion hole. Based on profile HMM-HMM comparisons, one further superfamily, which contains no protein structure data, was also assigned to the GGGX class and shows significant similarity in the oxyanion hole region. This class comprises the families abH1-abH6.

The Y-class consists of four superfamilies with known protein structures, where the oxyanion hole is not formed by a backbone amide, but by the hydroxyl group of a tyrosine side chain, which is strictly conserved within the superfamilies. This class comprises the families abH27-abH30.

To understand the stabilization of the structurally well conserved oxyanion hole architectures, sequence and structure data were compared and analyzed for these three classes.

For the GX-class representative structures of 11 superfamilies were investigated. Though the side chain of X is not involved in the catalytic mechanism it is well conserved within the superfamilies. The side chain of X is stabilized by one or several anchor residues, which are also well conserved within the families. If X is hydrophilic it is hydrogen bonded to hydrophilic anchor residues, while hydrophobic oxyanion hole residues bind to hydrophobic pockets. The conservation of the oxyanion hole residue and the anchor within the families indicated a modular organization of the first oxyanion hole residue and the anchor.

The stabilization of the oxyanion hole of the GGGX-class differs from the GX-class. The first oxyanion hole residue G is stabilized by interaction of the dipeptide GX with the side chain of the second oxyanion hole residue, which is a conserved alanine as C-terminal neighbour of the catalytic serine.

Since the oxyanion hole forming tyrosine of the Y-class can occupy two different positions the stabilization of the tyrosine containing loop can differ. For the Superfamilies abH27 and abH28 the tyrosine is located at the C-terminal end of the β-strand β3 of the central β-sheet, and thus is stabilized by this secondary structure element. For the superfamilies abH29 and abH30 the orientation of the tyrosine is stabilized by a preceding proline.

To proof the modularity of the first oxyanion hole residue and the anchor of the GX-class the hydrophobic module of the Lipase from *Burkholderia cepacia* (BCL) was exchanged for the hydrophilic module of the Lipase from *Rhizomucor miehei* (RML) in two steps. First the

oxyanion hole residue L17 was replaced by a threonine and second the anchor L167 was exchanged for an aspartate, glutamate, asparagine and glutamine like can be found in the homologous family of RML. While the replacement of the oxyanion hole residue resulted in a 200-fold decrease in activity against p-nitrophenylpalmitat the subsequent exchange of the anchor L167 with an asparagine resulted in the reactivation of the BCL. A 15-fold increase against p-nitrophenylpalmitat was observed compared to the single mutant, proofing the anchor concept.

Shape and physico-chemical properties of the scissile fatty acid binding sites of eight α/β hydrolases were analyzed and compared in order to understand the molecular basis of substrate specificity. Though all eight α/β hydrolases share the same fold and catalytic mechanism of ester hydrolysis, they differ in the substrate specificities for the acyl moiety. However α/β hydrolases can differ in geometry of their binding sites: lipases have a large, hydrophobic scissile fatty acid binding site, esterases like acetylcholinesterase and bromoperoxidase have a small acyl binding pocket, which fits exactly to their favorite substrates. The α/β hydrolases were devided into four subgroups: (1) lipases with a hydrophobic, crevice-like binding site located near the protein surface, (2) lipases with a funnel-like binding site, (3) lipases with a tunnel-like binding site and (4) α/β hydrolases with a cavity buried inside the protein. The length of the scissile fatty acid binding site varies considerably among lipases between 7.8 Å in cutinase and 22 Å in *Candida rugosa* lipase and *Rhizomucor miehei* lipase. Location and properties of the scissile fatty acid binding sites of all lipases of known structure were characterized. This model also identifies the residues which mediate chain length specificity and thus may guide protein engineering of lipases for changed chain length specificity. The model was supported by published experimental data on the chain length specificity profile of various lipases and on mutants of fungal lipases with changed fatty acid chain length specificity.

To utilize the LED for identifying new members of the α/β hydrolase fold family from genomic DNA or soil probes family specific PCR primers were designed. Using the method implemented in CODEHOP[11] a set of degenerated primers were inferred for the superfamily abH1 including short chain length specific lipases and carboxylesterases. The primers are based on two structurally relevant sequence motifs which are separated by around 100 amino acids and are specific for this family of α/β hydrolases. After validating the specificity of these primers using *Bacillus subtilis* as model organism, miscellaneous organisms were screened successfully for carboxylesterases.

The LED developed in this study has proven to be an efficient approach for a tight integration of vast amounts of protein family specific data. In combination with the methods provided by bioinformatics the LED is a useful tool for the systematic analysis of large protein families. Since the concept of the LED is not specific to α/β hydrolases only, this approach can be easily applied to further protein families to gain a deeper understanding of proteins' sequence-structure-function relationship.

1 Einleitung

1.1 Bestimmung der Funktion eines Proteins

Nachdem James B. Sumner 1926 die erste Aufreinigung und Kristallisation eines Proteins gelang, war es John Howard Northrop möglich, erstmals Pepsin zu isolieren und zu kristallisieren. Am Beispiel dieses Enzyms konnte Northrop 1930 schließlich zeigen, dass die katalytischen Eigenschaften des Pepsins eine intrinsische Eigenschaft dieses Proteins sind. Diese Experimente eröffneten ein neues Forschungsfeld, das noch heute eine der wichtigsten Aufgaben der modernen Molekularbiologie und Biochemie darstellt; die Entschlüsselung der Funktion von Proteinen. Aber selbst mit der Entwicklung moderner Hochdurchsatzmethoden und der biochemischen Charakterisierung unzähliger Proteine stellt diese Aufgabe immer noch einen langwierigen Prozess dar.

Parallel führten Techniken zur Sequenzierung ganzer Genome[12-14] und das neue Forschungsfeld der strukturellen Genomik[15,16] zu einer rasant ansteigenden Datenflut an biologischen Sequenzen und Proteinstrukturdaten. Für die Methoden der Bioinformatik stellt diese Datenmenge die beste Grundlage zur a priori Bestimmung der Funktion und Eigenschaften von Proteinen dar. Im Allgemeinen basieren diese Funktionsvorhersagen auf der Annahme, dass ähnliche Proteine evolutiv miteinander verwandt sind,[1] diese deshalb eine gemeinsames Faltungsmuster besitzen[2] und eine funktionelle Verwandtschaft besteht. In diesem Zusammenhang ist es wichtig eine Proteinsequenz im Kontext der Proteinfamilie, aus der die Sequenz hervorging, zu untersuchen.

Dieser deduktive Ansatz zur Bestimmung der Funktion eines Proteins ist jedoch nur erfolgversprechend für den Fall, dass gewisse methodische und konzeptionelle Bedingungen erfüllt sind: (1) da die Proteinfunktion aus den Sequenz-Struktur-Funktionsbeziehungen innerhalb einer Proteinfamilie abgeleitet wird, muss die Klassifikation der Proteine in Familien verlässlich sein. Falsch klassifizierte Proteine würden zu Inkonsistenzen innerhalb der Familie führen und die Datenanalyse negativ beeinflussen. (2) Die Untersuchung der Sequenz-Struktur-Funktionsbeziehungen setzt voraus, dass die hierfür nötigen Daten integriert analysiert werden können. Hierfür hat sich als die grundlegende Technik die vergleichende Sequenzanalyse über Multisequenz Alignments mehrfach bewährt.[17] (3) Diese Methodik basiert auf der Verarbeitung sehr großer Datenmengen. Die Heterogenität dieser Daten und das hohe Datenaufkommen verlangen jedoch nach einer Lösung zur Datenintegration, die einen effektiven Zugriff auf diese ermöglicht. Im folgenden werden diese drei Punkte genauer erörtert.

1.2 Proteinklassifikation

Mehrere Forschungsgruppen haben versucht Proteinsequenzen in homologe Familien und Superfamilien zu gruppieren. Die verschiedenen Ansätze unterscheiden sich im Grad der Automatisierung, dem umfassten Sequenzraum, der Fokussierung auf vollständige Proteine oder Proteindomänen und der zu Grunde liegenden Methodik. Allgemein lassen sich diese Ansätze in folgende Gruppen einteilen: Ansätze, die globale Proteinsequenzvergleiche zur Klassifikation heranziehen und andere, die wesentliche Merkmale einer Familie extrahieren und diese als Domänen oder Motive repräsentieren. Diese Deskriptoren können dann zur Untersuchung von Proteinsequenzen mit unbekannter Funktion verwendet werden. Die hohe Sensitivität dieser Sequenzanalysen haben gezeigt, dass diese Methode sehr wertvoll für die Bestimmung der Funktion unbekannter Proteine ist.

Die Sequenzmusterbibliothek PROSITE[18] stellt einen der ersten Ansätze dar, um familienspezifische Deskriptoren für funktionell relevante Proteinmotive zu beschreiben.[18] Diese Motive werden über reguläre Ausdrücke beschrieben, die nur die am höchsten konservierten Aminosäuren einbeziehen. Ein PROSITE Muster versucht die funktionell wichtigsten Aminosäuren zu identifizieren, wie z.B. das katalytische Zentrum eines Enzyms oder strukturell relevante Merkmale und beschreibt nicht eine Proteindomäne oder sogar ein vollständiges Protein. Sequenzmuster bewerten keine Ähnlichkeiten zwischen Muster und Sequenz, sondern beschreiben lediglich das Vorhandensein des Musters. Dies macht die PROSITE Muster zu unflexibel um stark divergierende Proteinfamilien zu beschreiben und viele der kürzeren Muster enthalten nicht genug Information um statistisch signifikante Treffer in den sehr großen Proteindatenbanken zu erzielen.

Ein verallgemeinerter Ansatz besteht darin, konservierte Sequenzbereiche aus Multisequenz Alignments zu extrahieren, deren Aminosäurevariationen innerhalb der Spalten des Alignments zu bestimmen und diese als *Position Specific Scoring Matrices* (PSSM) darzustellen.[19,20] Die konservierten Bereiche des Multisequenz Alignments, die zur Erstellung der PSSMs dienen, dürfen hierbei keine Lücken (*Gaps*) aufweisen, also Positionen, denen aufgrund von Insertionen in einer Sequenz keine Aminosäure zugewiesen werden kann. Blocks[21] und PRINTS[22] sind die bekanntesten Motivdatenbanken, die PSSMs nutzen um Proteinfamilien oder Proteindomänen zu repräsentieren.

Einen Schritt weiter geht die Definition von Profilen, die für jede Position eines Multisequenz Alignments einen *Score* für jede der 20 möglichen Aminosäuren ermitteln, wobei im Unterschied zu PSSMs auch *Gaps* über *Penalty Scores* berücksichtigt werden.[23] Die PROSITE Profilbibliothek, eine PROSITE Ergänzung, stellt solche Profile zur Verfügung.[18]

Eine verallgemeinerte Variante solcher Profile verwendet anstelle von *Score*s statistische Gewichtungen für jede Position des Profils, die die Wahrscheinlichkeit wiedergeben, dass eine der 20 Aminosäuren, eine Insertion oder eine Deletion vorkommen. Ein solches verallgemeinertes Profil wird als Hidden Markov Model (HMM) bezeichnet. Seit der Einführung dieser Methode zur Klassifikation von Proteinen[24] ist dessen Popularität stetig gestiegen. Die Datenbank Pfam[25] verwendet diese Methode um Proteindomänen zu klassifizieren.[25] Ähnlich wie die PROSITE Profile werden die Pfam HMMs iterativ überarbeitet. Beginnend mit einem Satz homologer Sequenzen werden weitere Familienmitglieder mit abnehmender Sequenzähnlichkeit in das HMM einbezogen. Aufgrund des hohen Informationsgehalts dieser Deskriptoren sind sie in der Lage auch weit entfernte Verwandte zu erkennen, die durch andere Methoden kaum zu finden sind.

Neben diesen Methoden zur Beschreibung von Familien über Deskriptoren der familienspezifischen Merkmale gibt es einige Ansätze, die Sequenz-Sequenz Vergleichsmethoden zur Klassifikation verwenden. PIR-ALN definiert drei verschiedene Klassifikationen,[26] die auf Sequenzidentität beruhen und manuell sorgfältig überarbeitet werden: funktionelle Domänen werden geclustert (PIR-D), Sequenzen mit mindestens 50% Identität in Familien gruppiert (PIR-F) und verwandte Familien werden in Superfamilien gesammelt (PIR-S). Hinter der Klassifikation von SBASE[27] steht ebenso ein spezialisierter Datenbank *curator*, der Sequenzvergleiche, Literaturzitate und einen *fuzzy clustering* Algorithmus für die Familienzuordnung nutzt. Es gibt jedoch auch verschiedene voll automatisierte Algorithmen die Sequenzähnlichkeit zur Einteilung von Proteinen nutzen. PROTOMAP verwendet eine Graphendarstellung der Proteinraums wobei Klassen von Proteinen aus der Verknüpfungsdichte abgeleitet werden.[28] Die Abstände zwischen den Proteinen werden über lokale Alignmentmethoden ermittelt.[29,30] SYSTERS verwendet einen *single-linkage* Clusteralgorithmus um Proteinfamilien zu definieren. Auch hier werden lokale Ähnlichkeit zwischen Proteinsequenzen verwendet, um deren Entfernungen zueinander abzuschätzen.

Während diese Ansätze die Klassifikation des gesamten Sequenzraums und der darin enthaltenen Proteine anstreben, existieren auch Datenbanken, die sich auf die Untersuchung und Klassifikationen einzelner Enzymklassen spezialisiert haben. Diese Datenbanken nutzen meistens den Ansatz des Sequenz-Sequenzvergleichs zur Klassifikation. Dieser familienspezifische Ansatz erlaubt im Gegensatz zu den globalen Sequenzklassifikationen eine detaillierte Untersuchung der Funktion individueller Proteine. ESTHER[31,32] wurde ursprünglich als Werkzeug zur Untersuchung von Cholinesterasen entworfen. Inzwischen auf

die Familie der homologen α/β-Hydrolasen ausgeweitet, analysiert ESTHER die Bedeutung dieser Familie für genetisch bedingte Erkrankungen des Menschen. CAZy[33] klassifiziert Enzyme, die die Umsetzung von Kohlenhydraten katalysieren, anhand von lokalen Sequenzähnlichkeiten und untersucht evolutive Prozesse innerhalb dieser Familie.

1.3 Die vergleichende Sequenzanalyse

Die vergleichende Sequenzanalyse ist eine der wichtigsten Methoden der modernen molekularen Biologie. Die ersten Analysen dieser Art gehen auf die Arbeiten Zuckerkandls und Paulings zurück.[1] Aber erst mit den Computer-Implementierungen von Dayhoff und Eck[34] und Needleman und Wunsch[35] konnten wohl definierte Sequenzvergleiche erzielt werden.

Das allen sequenzvergleichenden Methoden zugrunde liegende Problem ist das Auffinden des optimalen Alignments, also der besten Anordnung der sequentiellen Elemente zweier Proteinsequenzen (unter Einbeziehen eines Bewertungsschemas). In der einfachsten Form kann ein optimales Alignment als die minimale Anzahl von Mutationen definiert werden, die nötig sind um eine Sequenz in eine andere zu überführen. Die mathematische Lösung dieses Optimierungsproblems erfolgt über die dynamische Programmierung.[36] Drei verschiedene Ansätze zum Vergleich zweier Sequenzen können hierbei unterschieden werden: das optimale globale Alignment oder auch Needleman-Wunsch Alignment,[35] der Vergleich einer kürzeren Sequenz innerhalb einer deutlich längeren Sequenz und das lokale oder Smith-Waterman Alignment, das die ähnlichsten Sequenzsegmente bestimmt.[37] Neben diesen Optimierungsansätzen wurden auch effiziente Heuristiken zum Auffinden lokaler Ähnlichkeiten wie FASTA[29] und BLAST[30] entwickelt. Jedes Alignment wird über ein Maß für die Ähnlichkeit zwischen Paaren der Sequenzelemente bewertet. Solche Bewertungsschemata wie PAM250[34] oder BLOSUM62[38] werden als Matrizen logarithmischer Verhältnisse von Wahrscheinlichkeiten dargestellt, sogenannte *Scoring* Matrizen. Diese werden definiert als der Logarithmus der Wahrscheinlichkeit, dass zwei Aminosäuren an einer biologisch äquivalenten Position im Alignment vorkommen, im Verhältnis zur Wahrscheinlichkeit, dass diese beiden Aminosäuren in einem zufällig erzeugten Alignment gemeinsam auftreten.

$$Score_{ij} = \log\left\{\frac{P(a_i a_j)}{P(a_i)P(a_j)}\right\}$$

Allgemein kann man *Scoring* Matrizen als Darstellung der relativen Wahrscheinlichkeit für den Austausch einer Aminosäure gegen eine andere beschreiben. In die Berechnung eines Multisequenz Alignments muss auch ein *Gap Penalty* einbezogen werden, der die Wahrscheinlichkeit beschreibt, dass eine Deletion oder eine Insertionen in einer der Sequenzen auftritt.

Die Theorie der dynamischen Programmierung zum Auffinden eines optimalen Alignments zweier Sequenzen lässt sich in der Praxis nicht direkt auf das Multisequenz Alignment-Problem anwenden. Dies liegt daran, dass die Komplexität bei einer mittleren Sequenzlänge L und N Sequenzen $O(L^N)$ beträgt. Jedoch wurden eine Reihe von Heuristiken entwickelt, die das Multisequenz Alignment-Problem lösen. Meistens wird hierbei ein iterativer Ansatz verfolgt, der zuerst die beiden ähnlichsten Sequenzen zu einem Alignment vereint und dann die Sequenzen mit abnehmender Ähnlichkeit schrittweise dem Alignment hinzufügt. Sowohl die Handhabung von *Gaps* als auch die Bewertung der Ähnlichkeit, müssen für einen solchen iterativen Ansatz angepasst werden.

Der Umfang von Aminosäurevariationen an einer Position innerhalb eines Multisequenz Alignments hat sich als sehr hilfreiche Information erwiesen. So stimmen hoch konservierte Positionen sehr häufig mit Aminosäuren überein, die für die Funktion des Proteins entscheidend sind, z.B. das aktive Zentrum eines Enzyms, Disulfidbrücken oder zur Vermittlung der Substratspezifität. Deshalb ist es häufig möglich die Funktion oder die Struktur eines Proteins aus verwandten Familienmitgliedern abzuleiten.[39] So konnten aus Mustern für konservierte Sequenzbereiche eines Multisequenz Alignments funktionell relevante Aminosäuren vorhergesagt werden.[40] Konservierungsmuster wurden auch zur Bestimmung von Bindungsoberflächen verschiedener Proteinfamilien, mit gemeinsamen Faltungsmuster, eingesetzt.[41] Über den Grad der statistischen Kopplung zwischen Positionen in Multisequenz Alignments wurden Netzwerke interagierender Aminosäuren innerhalb einer Proteinfamilie identifiziert.[42] Mit Hilfe einer *Maximum Likelihood* Methode wurden Aminosäuren in Multisequenz Alignments lokalisiert, die eine korrelierte Evolution durchlaufen (Coevolution).[43] Für die Cytochrom C und Globin Familien[44,45] wurden konservierte Aminosäuren, die für die Funktion selbst nicht relevant sind, als für die Faltung relevant identifiziert. Die Sequenzvariabilitäten von Proteinen innerhalb eines gemeinsamen Faltungsmusters und ihrer Verbindung mit Faltungskeimen wurde ebenfalls anhand von Multisequenz Alignments untersucht.[46] Da Methoden zur Rekonstruktion phylogenetischer Bäume für homologe Proteinsequenzen auf Multisequenz Alignments aufbauen, kann die

Abstammung einer neuen Sequenz über einen solchen vergleichenden Ansatz bestimmt werden. Auch die vergleichende Sequenzanalyse ganzer Genome bedient sich der Multisequenz Alignments.[47]

1.4 Integration biologischer Daten

Gordon Moore sagte 1965 zutreffend voraus, dass sich die Anzahl der Transistoren, die sich auf einem Computerchip integrieren lassen, alle 18-24 Monate verdoppelt („Moore's law").[48] Neben der computergestützten Informationstechnologie ist das einzige Beispiel, das schneller Informationen anreichert, die Bestimmung biologischer Sequenzen.[49] Jedoch wächst nicht nur die Menge der biologischen Daten rasant an, auch die Diversität der Daten erhöht sich stetig. Waren zuerst nur primäre Daten wie biologische Sequenzen und Proteinstrukturen verfügbar, stehen inzwischen Datenbanken der zweiten und dritten Generation zur Verfügung, die komplexe biologische Systeme abbilden.

Die Integration dieser diversen Daten wird durch das Fehlen eines allgemeingültigen Konzepts zur Beschreibung biologischer Begriffe und Prozesse zusätzlich erschwert. Zwar gibt es Ansätze für kontrollierte Vokabulare wie z.B. Gene Ontology[50] (GO) zur Beschreibung der Rolle von Genprodukten, jedoch werden diese noch nicht allgemein verwendet und decken nicht alle Bereiche ab.

Des Weiteren bestehen technische Herausforderungen. Die existierenden biologischen Datenbanken verwenden unterschiedliche Datenbanksysteme und stellen keine standardisierten Schnittstelle für den Datenzugang bereit. Einige Datenbanken stellen textbasierte Dateien zur Verfügung, andere ermöglichen den Zugriff auf das zugrunde liegende Datenbanksystem oder repräsentieren den Inhalt der Datenbank in Form von Webseiten.

Zur Lösung dieser Probleme wurden drei Ansätze zur Integration heterogener Daten von verschiedenen Gruppen erprobt: Verknüpfung indizierter Daten, Federated Database Systeme und Data Warehousing.

1.4.1 Integration über verknüpfte, indizierte Daten

Diese Systeme verknüpfen Datenbanken, die auf einfachen Textdateien basieren, über den Hyperlinkmechanismus des World Wide Web (WWW). Beispiele hierfür sind SRS,[51] DBGET/LinkDB[52] und Entrez.[53] Obwohl diese Systeme keine wahre Datenintegration darstellen, erlauben sie dem Benutzer den Zugang über einen zentralen Einstiegspunkt und sind leicht zu implementieren. In der Praxis erwiesen sich diese Systeme als äußerst nützlich

und wurden schnell populär, was die Notwendigkeit zur Integration biologischer Daten widerspiegelt. Die Nachteile dieser Systeme sind, dass die Hyperlinks gepflegt werden müssen und anfällig für Fehler sind, die Datenanalyse nur über Browsing oder Stichwortsuchen erfolgen kann und ein großer Aufwand erforderlich ist, um die Daten für weitere Analysen aufzuarbeiten.[54]

1.4.2 Federated Database Systems

Ein Federated Database System (FDBS) verwirklicht die Integration heterogener Daten durch den Aufbau einer Umgebung, die die individuellen Datenbank als Teil eines größeren Systems erscheinen lässt. Die Quelldaten verbleiben hierbei in den individuellen Datenbanken. Der Zugriff wird jedoch über ein einheitliches Zugangssystem ermöglicht, das das FDBS bereitstellt.[55] Beispiele für solche Datenbanken übergreifende Systeme sind die Zugriffssprachen Kleisli und K2.[56] Das Zugangssystem ermittelt aus der Abfrage, welche Datenbanken einbezogen werden müssen um alle nötigen Daten für eine vollständige Antwort bereitzustellen, und generiert eine Reihe von Unterabfragen. Die Unterabfragen werden an die entsprechenden Datenbanken weitergeleitet und die resultierenden Ergebnisse durch das FDBS transformiert, integriert und an den Benutzer übergeben. Die Vorteile eines FDBS liegen darin, dass der Benutzer nur ein Zugriffssystem bedienen muss, das automatisch auf die heterogenen Daten zugreift. Die individuellen Datenbanken werden durch den Datenbank *curator* auf dem aktuellen Stand gehalten und der Benutzer erhält ein einzelnes, integriertes Ergebnis. Der Nachteil dieser Systeme liegt darin, dass die Antwortzeiten durch die verteilten Unterabfragen über das Internet lang sind und dadurch die interaktive Nutzung erschwert wird. Auch stellt der Ausfall einer integrierten Datenbank ein Problem dar, da in diesem Fall eine Anfrage nur unvollständig beantwortet werden kann.

1.4.3 Das Data Warehouse System

Dieser Ansatz vereint alle Daten der untersuchten Domäne in einem Datenbanksystem.[57] Hierzu muss zuerst ein einheitliches Datenmodell entwickelt werden, das sämtliche Daten halten kann, die von den relevanten Quelldatenbanken bereitgestellt werden. Um diese Datenintegration zu ermöglichen, muss eine Schnittstelle erstellt werden, die die Daten der Quelldatenbanken extrahiert, diese so transformiert, dass sie dem einheitliche Datenmodell des Data Warehouse entsprechen und schließlich die Daten im Warehouse ablegt. Die Vorteile dieser Art der Datenintegration bestehen darin, dass durch das einheitliche Datenmodell die heterogenen Daten semantisch abgeglichen werden und Probleme wie z.B. nicht standardisierten Nomenklaturen dadurch behoben werden. Das System ist unabhängig

von den Quelldatenbanken und dadurch resistent gegenüber einem Ausfall. Die Möglichkeit auch Daten aus Quellen zu integrieren, die nicht über eine Datenbank zugänglich sind, wie z.B. Publikationen oder eigenen Untersuchungen, erlaubt es Anfragen zu stellen, die durch die individuellen Quelldatenbanken nicht beantwortet werden können. Die Datenanalyse wird über ein einziges Zugangssystem ermöglicht und die Reaktionszeit des Systems ist deutlich schneller als der Zugriff auf individuelle Datenbanken über das Internet. Der große Nachteil dieses Ansatzes zur Datenintegration besteht darin, die Daten des Warehouse auf einem aktuellen Stand zu halten. Neue Daten werden permanent in die Quelldatenbanken eingespeist und müssen in das Data Warehouse übernommen werden, um einen aktuelles Abbild des Wissenstandes zu gewährleisten. Da sich die Quelldatenbanken jedoch weiterentwickeln, indem neue Datentypen, geänderte Nomenklaturen oder Relationen zwischen den Datentypen eingeführt werden, müssen die Schnittstellen zur Datentransformation ständig angepasst werden. Ebenso ist das dem Data Warehouse zugrunde liegende Datenmodell einem Entwicklungsprozess unterworfen, der der Integration neuer Informationen Rechnung tragen muss. Auch dies verlangt die unter Umständen zeitaufwendige Anpassung der Schnittstellen zur Datenintegration.

1.5 Datenbank Management Systeme

Die Notwendigkeit Daten zu speichern, diese zu organisieren und Anwendungen zur Analyse bereitzustellen führte bereits Ende der 60er Jahre zur Entwicklung von Databank Management Systemen (DBMS). Zentraler Punkt eines jeden DBMS ist das konzeptionelle Schema. Dieses Schema enthält die Beschreibung aller für die untersuchte Domäne relevanten Datenstrukturen und ist stark vom Datenbankentwurf und dem benutzen Datenmodell (relational, objektorientiert, hierarchisch, XML) abhängig. Zum konzeptionellen Schema gehören neben allen relevanten Beschreibungen für die Datenhaltung insbesondere auch alle Beschreibungen von Integritätsregeln. Dabei ist es das Ziel, dass das DBMS, unabhängig von Anwendungsprogrammen, im Betrieb dafür sorgt, dass diese Integritätsregeln eingehalten werden. Andernfalls würden viele Regeln mehrfach im Programmcode repräsentiert sein mit der Folge von möglichen Widersprüchen in verschiedenen Anwendungen und einer verschlechterten Wartbarkeit. Dies könnte zu nicht überschaubaren Auswirkungen auf die gesamte Datenbank führen. Ein DBMS stellt auch eine Schnittstelle zur Datenbank und deren Datenschema bereit. Diese Schnittstelle erlaubt die Manipulation der gespeicherten Daten und regelt die Zugriffsrechte.

1.5.1 Hierarchische Modelle

Die ersten DBMS der 60er Jahre verwendeten eine hierarchische Beschreibung des konzeptionellen Schemas (HDBMS). Die hierarchischen Modellierung eines Datenbankschemas beruht auf einem streng definierten Baum bestehend aus Datenkonten. Jeder Knoten kann Daten sowie einen Satz Unterknoten besitzen. Die Anzahl der Unterknoten kann zwischen den Geschwisterknoten einer Ebene variieren, aber der Typ aller „Cousins“ ist identisch. Der Datenzugriff in einem hierarchisch organisierten Schema folgt einem konkreten Pfad durch die Datenstruktur zum gewünschten Datensatz, der so vom DBMS schnell bestimmt werden kann. Der Nachteil dieser Modellierung liegt darin, dass Abfragen, für die das hierarchische Datenmodell nicht vorgesehen ist, zu komplexen Kombinationen verschiedener Pfade führen kann.

1.5.2 Objektorientierte Modelle

In den 80er Jahren wurden die ersten DBMS vorgestellt, die auf einer objektorientierten Modellierung der Datenbankschemata basierten (OODBMS). Diese wurden eingeführt um Anwendungsgebiete wie CAD, Dokumentenverwaltung oder Multimediaanwendungen zu unterstützen, die durch andere Datenmodellierungen nur unzureichend beschrieben werden konnten. Die Basiselemente der objektorientierte Modellierung eines Datenbankschemas sind Objekte.[58] Ein Objekt befindet sich zu jedem Zeitpunkt in einem definierten Zustand, der sich jedoch im Laufe der Zeit ändern kann. Der Zustand eines Objekts wird durch Werte beschrieben, die seine Eigenschaften aufweisen und durch die Beziehungen, die er zu anderen Objekten hat. Objekte besitzen auch Verhaltensregeln (Methoden), die beschreiben wie das Objekt auf Änderungen reagiert.

Die Beziehungen zwischen den Objekten sind hierarchisch organisiert weshalb die objektorientierte Modellierung in gewisser Weise mit der hierarchischen Modellierung verglichen werden kann. In diesem Sinne sind Objekte den inneren Knoten in einem hierarchischen Modell ähnlich, die ebenfalls einen Satz an Unterknoten besitzen. Der große Unterschied liegt jedoch in der Tatsache, dass Objekte heterogen sein können. Die Heterogenität eines Objektes bedeutet, dass jedes Objekt nur den Satz an Attributen besitzt, die benötigt werden um es zu beschreiben, im Gegensatz zu den hierarchisch organisierten Knoten.

1.5.3 XML Modelle

XML hat sich im Laufe der letzten Jahre zu einem der wichtigsten Formate zum Austausch von Daten etabliert. Im Rahmen dieser Entwicklung kam die Notwendigkeit auf, diese XML Dokumente effizient zu verwalten. Zur Lösung dieses Problems wurden DBMS entwickelt die auf der Modellierung des Datenbankschemas in XML basieren. Diese nativen XML DBMS (XDBMS) führen sämtliche Datenmanipulationen wie Einfügen, Löschen oder Sichten direkt am XML Objekt und gemäß dessen Struktur durch. Die Modellierung eines Datenbankschemas über XML ähnelt am ehesten dem der OODBMS. Ein XML Schema besteht ebenfalls aus Knoten, die heterogene Daten enthalten können. XDBMS eignen sich besonders für die Verwaltung dokumentorientierter XML Daten wie z.B. Formulare. Diese auf einen Benutzer zugeschnittenen Dokumente enthalten häufig gemischten Inhalt und sind weniger strukturiert als Daten XML Dokumente. Der Nachteil der XDBMS liegt wie bei den HDBMS darin, das strukturfremde Anfragen nur schlecht realisierbar sind.

1.5.4 Relationale Modelle

Relationen wurden 1970 erstmals als mögliche Grundlage einer mathematisch funktionierenden Datenbanktheorie untersucht.[59] In den folgenden zehn Jahren führten diese Grundlagen zur Entwicklung der relationalen DBMS (RDBMS) und der sogenannten Relationenalgebra, die das Werkzeug zur Datenmanipulation in RDBMS darstellt.[60] Ein relational modelliertes Datenbankschema besteht aus einem Satz von Relationen, die vereinfacht als Tabellen angesehen werden können. Diese Tabellen beschreiben eigenständige Einheiten, sogenannte Entitäten, der untersuchten Domäne. Eine Entität der Domäne Proteinfamilie wäre z.B. ein Protein, dessen Struktur oder die zugehörige Familie. Jede Tabelle besitzt eine feste Anzahl von Spalten, den Attributen der Relation. Diese Attribute beschreiben die Eigenschaften der zugehörigen Entität. Eine unbestimmte Anzahl von Zeilen (Tupel) stellen die gespeicherten Datensätze in den Tabellen dar. Die so definierten Relationen können im Datenmodell miteinander verknüpft werden und Beziehungen zwischen Entitäten wiedergeben.

Um einzelne Tupel einer Relation zu manipulieren, müssen alle Tupel eindeutig identifiziert werden können. Hierzu werden Primärschlüssel definiert. Ein Primärschlüssel einer Relation ist durch eine Menge von Attributen festgelegt. Bei einem Primärschlüssel, der aus einem Attribut besteht, darf in einer Relation zu einem bestimmten Zeitpunkt jeder Wert für das Attribut nur einmal auftreten. Bei Schlüsseln, die aus mehreren Attributen bestehen, dürfen

die entsprechenden Kombinationen der Attribute nur jeweils einmal auftreten. Diese Bedingung wird Entitätsintegrität genannt.
Bei Relationen, die konkrete Entitäten (z.B. Protein, Proteinstruktur) bezeichnen, erfüllen die natürlichen Attribute der Entität häufig nicht die Bedingung der Entitätsintegrität. In diesen Fällen werden üblicherweise künstliche Primärschlüssel eingefügt, wie z.B. ein sich selbständig erhöhender Zähler.
Die Verknüpfungen zwischen Relationen erfolgen über Werte. Diese Werte, die Beziehungen zwischen den verschiedenen Relationen beschreiben, werden als Fremdschlüssel bezeichnet. Die Werte dieser Fremdschlüssel beziehen sich jeweils auf die Werte des Primärschlüssels einer anderen Relation. Für diese Beziehungen wird allgemein die referenzielle Integrität gefordert, was bedeutet, dass der Wert des Fremdschlüssels in der abhängigen Relation auf jeden Fall als Primärschlüsselwert in der verknüpften Relation enthalten sein muss.
Für die Implementierung der Relationsalgebra zur Datenmanipulation in relationalen Datenbanken, hat sich SQL (Structured Query Language) als Standard etabliert. Diese 1987 erschienene Abfragesprache stellt die Schnittstelle zwischen der relationalen Datenbank und dem Anwendungsprogramm dar. Unter der Aufsicht der ISO und der IEC wurde im Dezember 2003 die fünfte Version dieses Standards, SQL:2003, veröffentlicht. Nicht zuletzt ist die Eigenschaft von SQL komplexe Abfragen effizient formulieren zu können der Grund für die vorherrschende Stellung der RDBMS unter den Datenbankanwendungen. Die kontrollierte Weiterentwicklung dieser Sprache half auch neue Anwendungsfelder zu unterstützen. So wurden in der Version SQL:1999 objektorientierte Erweiterungen implementiert und die aktuelle Version SQL:2003 enthält nun auch Funktionen zur direkten Manipulation von XML Objekten und externen Daten.

1.6 α/β-Hydrolasen

Die α/β-Hydrolasen sind eine der größten Familien von strukturell verwandten Enzymen, die ein breites Spektrum an chemischen Reaktionen katalysieren und dabei eine Vielzahl verschiedenster Substrate umsetzen.[6] Mitglieder dieser Familie sind unter anderem Esterasen (E.C. 3.1.1), Acetylcholinesterasen (E.C. 3.1.1.7), Cutinasen (E.C. 3.1.1.74), Carboxylesterasen (E.C. 3.1.1.1), Arylesterasen (E.C. 3.1.1.2), Phospholipasen A1 (E.C. 3.1.1.32), Cholinesterasen (E.C. 3.1.1.8), Juvenile Hormon Esterasen (E.C. 3.1.1.59), Thioesterasen (E.C. 3.1.2.14), Epoxidhydrolasen (E.C. 3.3.2.3), Hydroxynitrillyasen (E.C.

4.1.2.39), Lysophospholipasen (E.C. 3.1.1.5), Acyltransferasen (E.C. 2.3.1), Dipeptidyl-peptidasen (E.C. 3.4.14.5) und Non-heme Peroxidasen (E.C. 1.11.1.10).

Der dreidimensionale Fold eines Proteins stellt das architektonische Gerüst für die enzymatische Funktion und die Substratspezifität dar. Es gibt zwei Modelle für die Evolution der Enzymstruktur und Funktion. In der konvergierenden Evolution entwickeln sich Proteinstrukturen oder Folds unabhängig voneinander um ähnliche Funktionen durchzuführen. Ein Beispiel hierfür ist die Familie der Serinproteasen, die mit Trypsin[61] und Subtilisin[62] Vertreter mit völlig unterschiedlichen Folds enthält, beide aber effektive Katalysatoren für die Hydrolyse von Peptiden darstellen. Die häufiger auftretende divergierende Evolution beschreibt die Entwicklung von Proteinen mit unterschiedlichen Funktionen, ausgehend von einem gemeinsamen Vorgänger. Es ist bekannt, dass eine Vielzahl verschiedener Proteinsequenzen einen ähnlichen Fold repräsentieren können. Es scheint, dass in der Natur stabile Faltungsmuster dazu gewählt wurden, um als Vorlage für die molekulare Anpassung der Enzymfunktion zu dienen. Innerhalb eines Folds schließt dieser Prozess den Austausch von Aminosäuren sowie Insertionen oder Deletionen von Loops ein, die zu Enzymen mit verschiedenen katalytischen Funktionen und nur wenigen konservierten Sequenzbereichen führen können. Der α/β-Hydrolase Fold stellt solch ein Beispiel dar.[5]

1.6.1 Der α/β-Hydrolase Fold

Der α/β-Hydrolase Fold gehört zum *doubly wound* α/β-Superfold[3,4] und stellt einen von 121 verschiedenen Folds in der SCOP Klassifikation von überwiegend parallelen α/β Strukturen dar.[63] Dieser Fold wurde erstmals nach der Untersuchung von Proteinstrukturen für fünf Enzyme beschrieben:[5] Dienlacton Hydrolase,[64] Haloalkan Dehalogenase,[65] Weizen Serincarboxypeptidase II,[66] Acetylcholinesterase[67] und die Lipase aus *Geotrichum candidum*.[68] Obwohl alle dieser fünf Enzyme Hydrolysen katalysieren, unterscheiden sich die hydrolysierten Substrate deutlich voneinander. Nur Acetylcholinesterase und die Lipase aus *G. candidum* teilen eine ausreichend hohe Sequenzähnlichkeit um eine ähnliche Struktur vorherzusagen.[69] Aufgrund von Strukturüberlagerungen dieser fünf Strukturen wurden die konservierten Elemente dieses neuen Folds beschrieben.[5]

Das aktive Zentrum jedes dieser fünf Enzyme besteht aus einer katalytischen Triade ähnlich zu der, die zuvor für Serinproteasen beobachtet wurde. Jedoch können die α/β-Hydrolasen von den Familien der Proteasen (Trypsin, Papain und Subtilisin) anhand der Anordnung der katalytisch aktiven Aminosäuren unterschieden werden. In α/β-Hydrolasen ist die Abfolge der Aminosäuren der katalytischen Triade immer Nucleophil, Säure, Histidin im Vergleich zu

His-Asp-Ser für die Trypsin Familie, Cys-His-Asn für die Papain Familie und Asp-His-Ser für die Subtilisin Familie.[5] Für eine virale Cysteinprotease wurde eine His-Glu-Cys Triade beobachtet und könnte eine fünfte Enzymfamilie mit einer stabilen katalytischen Triade darstellen.[70] Für die α/β-Hydrolasen sind die Aminosäuren der katalytischen Triade variabler als im Vergleich zu den Proteasefamilien. Serin, Aspartat und Cystein wurden alle als Nucleophil in α/β-Hydrolasen identifiziert. Serin ist jedoch zwingend notwendig für Serinproteasen[71] und Cystein für die Papain Familie um eine signifikante Aktivität zu erhalten.[72] Des Weiteren wurde für einzelne Vertreter der α/β-Hydrolasen erstmals ein Glutamat anstelle eines Aspartats als katalytische Säure beobachtet.[67,68]

Der α/β-Hydrolase Fold besitzt einen Kern, der aus einem zentralen, überwiegend parallelen β-Faltblatt besteht (Abbildung 1). Nach der Nomenklatur von Richardson[3] folgt die kanonische Verknüpfung der acht β-Strands der Topologie +1, +2, -1x, +2x, $(+1x)_3$. Das β-Faltblatt besitzt eine linkshändige Verdrillung, so dass der erste und letzte β-Strand sich in einem Winkel von etwa 90° kreuzen.[5] Die Verdrillung des β-Faltblatts unterscheidet sich dabei deutlich zwischen den verschiedenen Enzymen. Helices sind auf beiden Seiten des β-Faltblatts angelagert. Die Helices A und F befinden sich auf der konkaven Seite des β-Faltblattes und Helices B bis E liegen auf der konvexen Seite. Die Positionen der Helices im Raum variieren stark zwischen den Enzymen, trotz der gemeinsamen Topologie. Die Überlagerung des β-Faltblatts der α/β-Hydrolasen führt in den meisten Fällen zu einer schlechten Übereinstimmung der α-Helixpositionen im Raum. Die α-Helix C, die einen Teil des *nucleophilic elbow*s formt, stellt hierbei eine Ausnahme dar. Die zentrale Lage dieser α-Helix und deren Beitrag für die Ausrichtung des katalytischen Nucleophils setzten voraus, dass deren Position innerhalb der α/β-Hydrolasen stark konserviert sind.

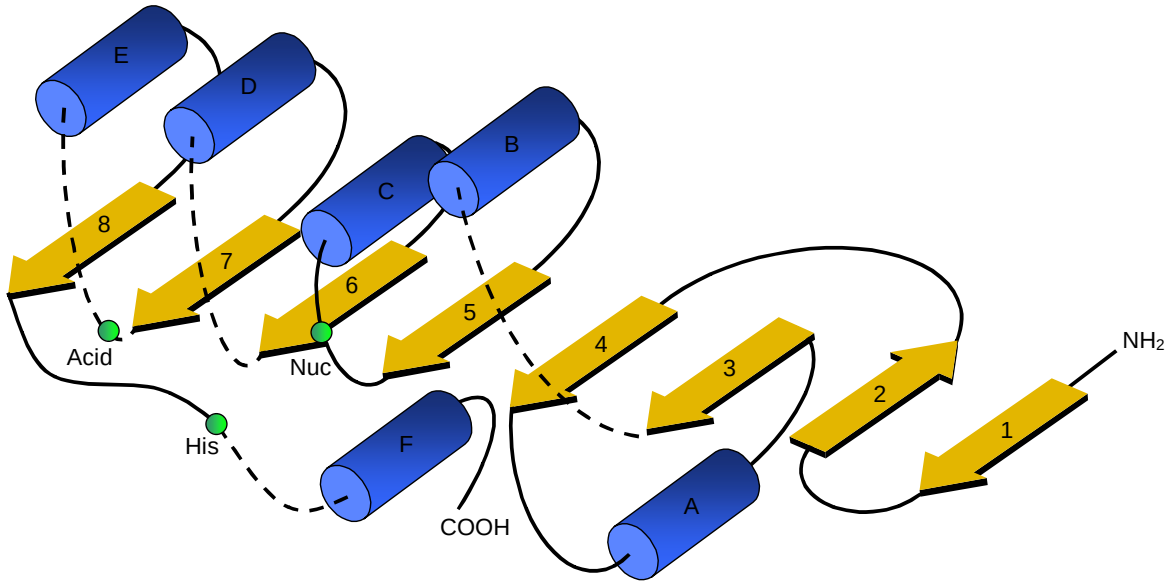


Abbildung 1: Schematische Darstellung des α/β-Hydrolase Folds. α-Helices sind als blaue Zylinder (A-F) dargestellt, β-Strands als gelbe Pfeile (1-8). Die gestrichelten Linien deuten variable Bereiche an. Der *Nucleophile Elbow* ist zwischen β-Strand β5 und α-Helix αC angedeutet. Die Positionen der Aminosäuren der katalytischen Triade sind als grüne Punkte dargestellt.[5]

1.6.2 Der *Nucleophilc Elbow*

Die katalytisch aktiven Aminosäuren sind an C-terminalen Enden der β-Strands lokalisiert. Das katalytische Nucleophil befindet sich in dem hoch konservierten Pentapeptid G-X-S-X-G. Dieses häufig als Consensus Sequenz beschriebene Motiv ist jedoch nicht in allen α/β-Hydrolasen vollständig erhalten.[5,73] Wie bereits beschrieben kann das katalytische Nucleophil ein Serin, Aspartat oder Cystein sein und die Glycine an den Positionen Nu-2 und Nu+2 können durch andere, kleine Aminosäuren wie Alanin oder Serin ausgetauscht sein.[74] Das Pentapeptid um das katalytische Nucleophil bildet eine enge Schleife in einem β-Strand-Turn-α-Helix Motiv. Dies zwingt das Nucleophil dazu, ungewöhnliche Backbone Torsionswinkel ϕ und ψ von –50° und –130° einzunehmen. Der in dieses Motiv einbezogene β-Strand liegt im allgemeinen nahe dem Zentrum des β-Faltblatts und ist β5 im kanonischen Fold. Die α-Helix C stellt die Helix in diesem Strukturmotiv dar. Dieser *nucleophilic elbow* ist das am besten konservierte Element des α/β-Hydrolase Folds. Die sterischen Einschränkungen zur Ausbildung des Turns bedingen kleine oder keine Seitenketten an den Positionen Nu-2 und Nu+2 und in den meisten Fällen sind diese Positionen daher durch

Glycine besetzt. Die Aminosäure an Position Nu+3 muss ebenfalls eine kleine Seitenkette besitzen um sterische Wechselwirkungen mit β-Strand β5 zu vermeiden.

1.6.3 Die katalytische Säure und das Histidin

Die katalytische Säure ist in einem Loop nach dem β-Strand β7 lokalisiert, der einen *reverse* Turn darstellt. Die H-Brückenstabilisierung der Säureseitenkette variiert abhängig davon, ob es sich um ein Aspartat oder ein Glutamat handelt.[5,75] In beiden Fällen nimmt ein Sauerstoffatom der Carboxylgruppe eine ähnliche Position ein und bildet eine H-Brücke zu einem Stickstoffatom des Imidazolrings des Histidins aus. Das katalytische Histidin befindet sich in einem Loop nach dem β-Strand β8. Länge und Konformation dieses Loops sind variabel.

1.7 Lipasen

Triacylglycerol Lipasen sind ubiquitäre, wasserlösliche Enzyme, die den hydrolytischen Abbau neutraler Lipide zu Fettsäuren und Alkoholen katalysieren. Diese Enzyme sind effiziente Katalysatoren für die Spaltung der Esterbindungen von Mono-, Di- und Triglyceriden in wässrigen Emulsionen. Sie stellen eine diverse Gruppe an Enzymen dar, die unterschiedlichste enzymatische Eigenschaften und Substratspezifitäten besitzen.[76] Im Allgemeinen akzeptieren diese Enzyme ein breites Spektrum an natürlichen und synthetischen Estern als Substrate, weshalb sie häufig in technischen Anwendungen der Biotechnologie eingesetzt werden.[77]

Die meisten mikrobiellen Lipasen sind extrazelluläre Enzyme, die vom Mikroorganismus sekretiert werden, um Lipide zu Produkten zu hydrolysieren, die dann vom Organismus für dessen Wachstum verwendet werden können. Diese Enzyme mit einer molekularen Masse von 19-60 kDa zeigen jedoch nur eine geringe lipolytische Aktivität in Gegenwart von geringen Konzentration des Substrats.

1.7.1 Grenzflächenaktivierung

1958 untersuchten Sarda und Desnuelle[78] die lipolytische Aktivität der pankreatischen Lipase und beschrieben das Phänomen der Grenzflächenaktivierung. Dieses Phänomen beschreibt die Tatsache, dass die Aktivität der Lipase in Gegenwart von Lipiden gesteigert wird. Desnuelle et al. postulierten, dass Lipasen durch die Adsorption an der Wasser-Lipid Grenzfläche aktiviert werden und dass diese Aktivierung mit einer Konformationsänderung des Enzyms in Verbindung steht.[78,79] Diese Grenzflächenaktivierung wurde später für eine Vielzahl weiterer

Lipasen beobachtet und wurde zum charakteristischen Merkmal um Lipasen von Esterasen zu unterscheiden. Jedoch zeigten die Untersuchungen kürzlich entdeckter Lipasen, die trotz deutlicher Sequenzähnlichkeit zu bekannten Lipasen keine Grenzflächenaktivierung zeigten,[80-82] dass eine Klassifikation nur anhand der Grenzflächenaktivierung nicht möglich ist.

1.7.2 Der α/β-Hydrolase Fold der Lipasen

Seit der Beschreibung des α/β-Hydrolase Folds anhand von fünf Proteinstrukturen[5] wurden für eine Vielzahl weiterer Vertreter dieser Familie die Proteinstrukturen veröffentlicht, darunter auch mehrere Lipasen. Mit wenigen Ausnahmen gehören die strukturell bestimmten Lipasen zum α/β-Hydrolase Fold und weisen dessen Merkmale auf. Sie besitzen ein zentrales, überwiegend paralleles β-Faltblatt und das konservierte Pentapeptid G-X-S-X-G ist im *nucleophilic elbow* lokalisiert. All diese Lipasen mit bekannter Proteinstruktur entsprechen somit dem ursprünglich beschriebenen Faltungsmuster der α/β-Hydrolasen, jedoch existieren auch Variationen.

1.7.3 Zwei Klassen von Konformere: geschlossen und offen

Ein wichtiges Merkmal von Enzymen ist deren Form der Substratbindungstasche und deren Zugänglichkeit für das Substrat. Proteinstrukturen der meisten Enzyme zeigen, dass deren aktive Zentren sich auf der Proteinoberfläche befinden und für das Lösungsmittel zugänglich sind. Dies trifft auf Lipasen nicht immer zu. Die bisher bestimmten Lipasestrukturen können in zwei Kategorien eingeteilt werden: solche mit einem für das Lösungsmittel zugänglichen aktiven Zentrum (offene Form) und solche mit einem nicht zugänglichen aktiven Zentrum (geschlossene Form). Für einige Lipasen wurden beide Konformationen experimentell beobachtet. Die Tatsache, dass beide Formen nicht nur innerhalb einer homologen Familie auftreten sondern über die Familiengrenzen hinaus zu beobachten sind spricht dafür, dass diese Zustände für die Funktion dieser Enzyme wichtig sind.

1.7.4 Die geschlossene (inaktive) Konformation

Die geschlossene Konformation wird als die im wässrigen Medium überwiegend vorliegende Spezies vermutet, was durch die geringe lipolytische Aktivität der meisten Lipasen in Abwesenheit einer Grenzfläche untermauert wird. Gleichzeitig zeigt aber das Vorhandensein einer geringen Aktivität auch im wässrigen Medium, dass möglicherweise andere Konformationen dem Substrat den Zugang zum aktiven Zentrum zumindest vorübergehend erlauben. Die geschlossene Konformation ist über ein nicht besetztes aktives Zentrum

charakterisiert, das vom Lösungsmittel durch ein oder mehrere Loops abgeschirmt ist. Diese Loops bilden das sogenannte Lid. Die Proteinoberfläche der geschlossen Form der Lipasen ist wie für wasserlösliche Proteine typisch überwiegend hydrophil.

1.7.5 Die offene (aktive) Konformation

Die Eigenschaften der Lipaseoberfläche in der offenen Konformation unterscheiden sich deutlich zu denen der geschlossenen Konformation. Die Bewegung des Lids öffnet nicht nur den Zugang zum aktiven Zentrum, sondern erzeugt auch eine große hydrophobe Oberfläche. In der Lipase aus *Rhizomucor miehei* führt die Konformationsänderung während der Aktivierung zu einer Vergrößerung der hydrophoben Oberfläche um etwa 700 Å^2.[83] Gleichzeitig wird ein Teil der zuvor exponierten hydrophilen Oberfläche verdeckt. Die Abnahme der hydrophilen Oberfläche beträgt hierbei etwa 450 Å^2. Eine deutlich größere hydrophobe Oberfläche von mehr als 1000 Å^2 wird sogar im Fall der Lipase aus *Candida rugosa* generiert.[84] In gleicher Weise wurde für die pankreatische Lipase eine Zunahme der hydrophoben Oberfläche um das aktive Zentrum beschrieben.[85]

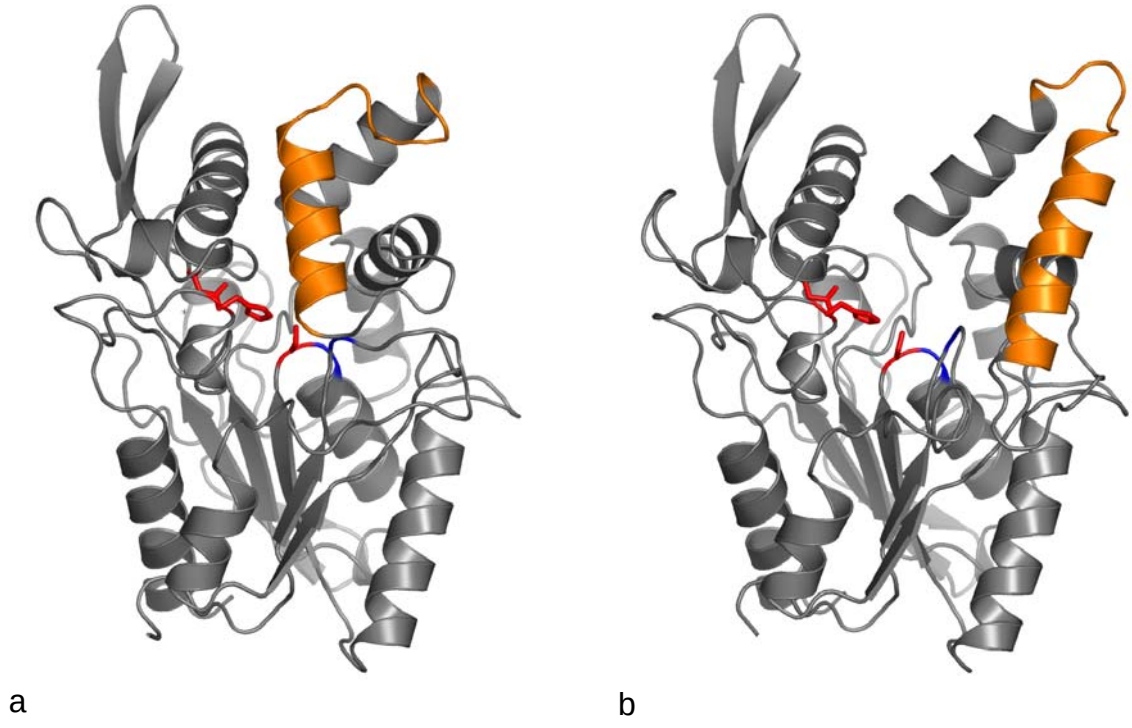


Abbildung 2: Darstellung der geschlossen (inaktiven) Form der Lipase aus *Burkholderia glumae* (a) und der offenen (aktiven) Form der homologen Lipase aus *Burkholderia cepacia* (b). Das Lid ist in orange dargestellt, die katalytische Triade in rot, die das *oxyanion hole* bildenden Aminosäuren in blau.

1.7.6 Das *Oxyanion Hole*

Eine wichtige Komponente des katalytischen Mechanismus der Hydrolasen ist das erstmals für Serinproteasen postulierte *oxyanion hole*.[61,86] Dieses *oxyanion hole* beschreibt eine Umgebung von zwei wohl definierten Wasserstoffbrückendonoren (im Allgemeinen Amidgruppen des Backbones). Diese sind für die Stabilisierung des während der Reaktion auftretenden Übergangszustandes zuständig, in dem das Carbonylsauerstoffatom des Substrats eine partiell negative Ladung trägt. In Serinproteasen liegt dieses *oxyanion hole* auch im substratfreien Zustand in einer funktionellen Form vor, dies ist für viele Lipasen nicht der Fall. Für die meisten Lipasen ist zumindest eine der beiden *oxyanion hole* bildenden Aminosäuren auf einem der während der Aktivierung beweglichen Loops lokalisiert. Dies führt dazu, dass sich das *oxyanion hole* für diese Lipasen nur in der offenen Form in einem funktionellen Zustand befindet.

1.7.7 Reaktionsmechanismus der Lipasen

Der katalytische Mechanismus[87] der Lipasen ist in Abbildung 3 skizziert: das katalytisch aktive Histidin abstrahiert vom Serin ein Proton, wodurch die Nucleophilie des Serinrests ansteigt. Dieser kann nun an dem Carbonylkohlenstoff eines Substratesters angreifen, der bereits in der Bindungstasche über Van-der-Waals Wechselwirkungen gebunden vorliegt. Es bildet sich eine tetrahedrale Übergangsstufe, aus der sich im zweiten Schritt durch Austritt eines Alkoholats ein Acyl-Enzymkomplex bilden kann. Dieser wird durch einen Hydrolyseschritt, wiederum über eine tetrahedrale Stufe, in die Carbonsäure und das freie Enzym gespalten.

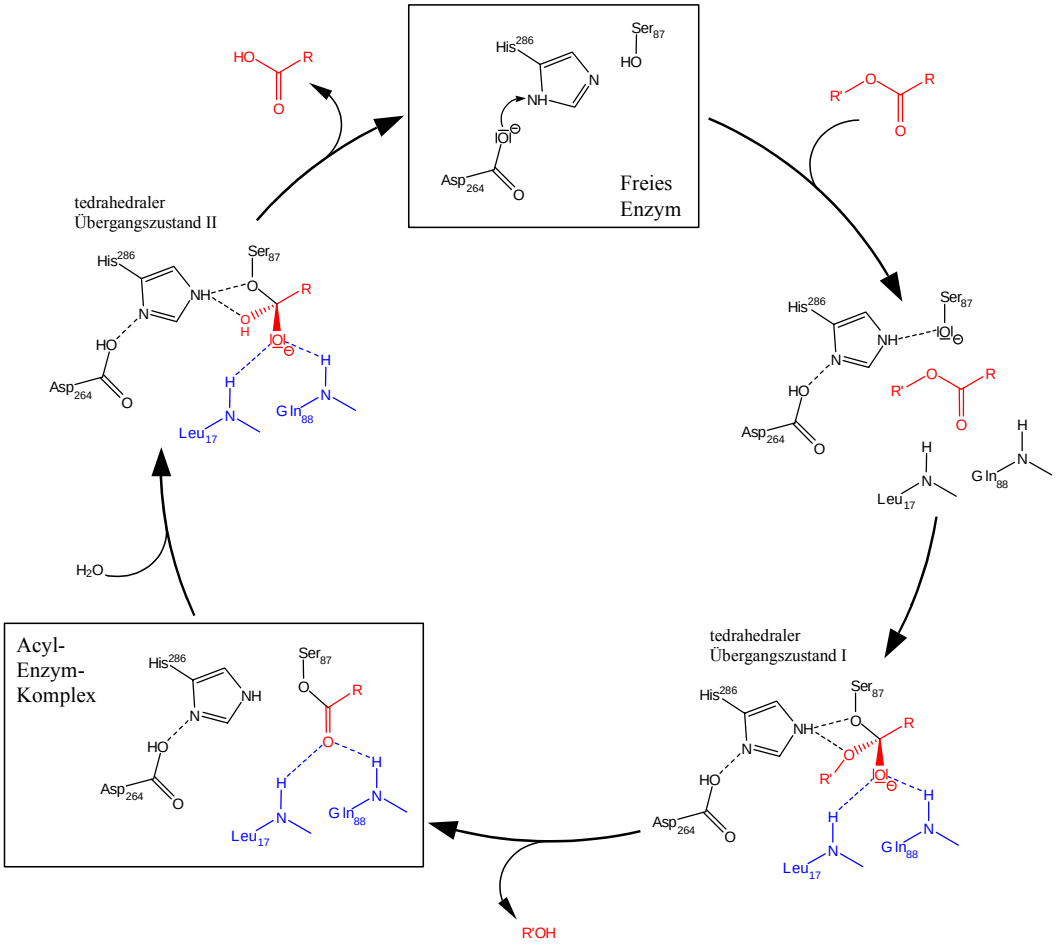


Abbildung 3: Schematische Darstellung des Reaktionsmechanismus einer Lipase am Beispiel der Hydrolyse eines Esters. Das Substrat wurde rot dargestellt, die Aminosäuren des *oxyanion holes* blau. Die Bezifferung der katalytisch aktiven Reste entspricht der Lipase aus *Burkholderia cepacia*.

2 Zielsetzung

Im Rahmen dieser Arbeit sollte mit den Methoden der Bioinformatik die Familie der α/β-Hydrolasen systematisch analysiert werden. Diese Analyse sollte zu einem besseren Verständins der Sequenz-Struktur-Funktionsbeziehung dieser diversen Proteinfamilie beitragen. Voraussetzung für diese Art der Untersuchung war die Entwicklung eines *Data Warehouse* Systems zur automatisierten Integration der heterogenen Sequenz- und Strukturdaten sowie funktionell relevanter Informationen. Mit Hilfe dieses Systems sollte eine verlässliche Klassifikation der α/β-Hydrolasen über Sequenz- und Strukturvergleiche etabliert werden, die sich als Grundlage der systematischen Analyse eignete.

Für die Isolierung neuer α/β-Hydrolasen aus Bodenproben oder genomischer DNA sollten Sequenzmotive identifiziert werden, die zur Konstruktion familienspezifischer Hybridprimer verwendetet werden können.

Eine schematische Modellierung der Fettsäurebindungstaschen verschiedener Lipasen und Esterasen sollte die unterschiedlichen Substratspezifitäten beschreiben. Die Modelle sollten auch durch Mutationen veränderte Substratspezifitäten erklären und zur Vorhersage von gezielten Änderungen beitragen.

Die Konservierungsanalyse der Aminosäuren innerhalb der Familien der α/β-Hydrolasen sollte in Kombination mit Strukturuntersuchungen die Funktionsweise des *oxyanion holes* und dessen Stabilisierung klären sowie das für den α/β-Hydrolase Fold charakteristische Konservierungsmuster beschreiben.

3 Ergebnisse

3.1 Relationales Datenmodell der LED

Für das Data Warehouse System wurde ein relationales Datenmodell entwickelt, das geeignet ist um die in dieser Arbeit untersuchten Beziehungen zwischen der Proteinsequenz, der Proteinstruktur und der Funktion des Proteins verschiedener Protein Familien zu analysieren (Abbildung 4). Um die für diese Aufgabe benötigten Daten zu integrieren wurde das Datenmodell in 3 Hauptbereiche eingeteilt: (1) Entitäten, die der Definition des Begriffs Protein gewidmet sind, (2) Entitäten, die die Proteinsequenz und deren Eigenschaften beschreiben und (3) Entitäten, die die Proteinstrukturen definieren.

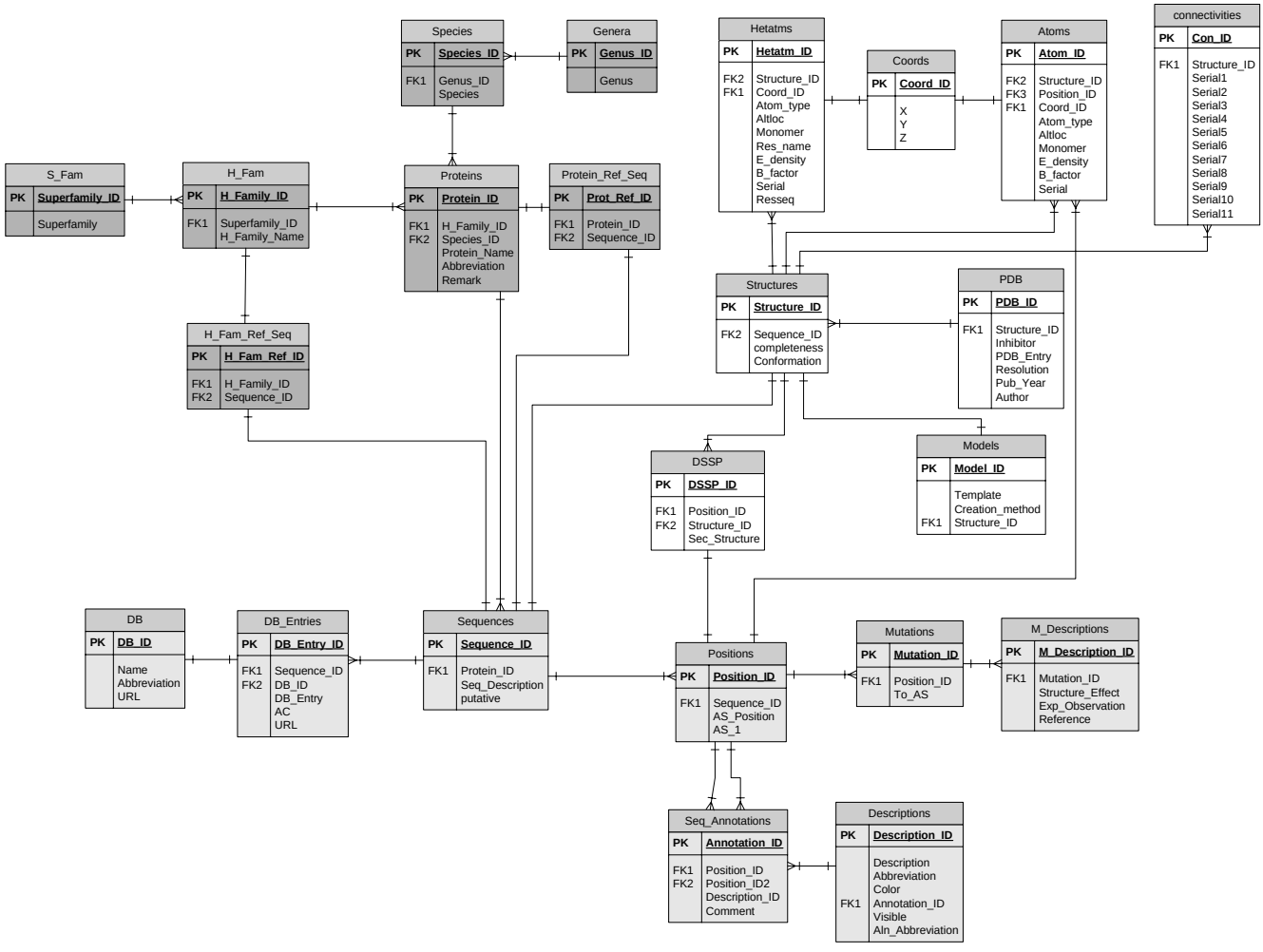


Abbildung 4: Relationales Datenmodel der LED. Die Entitäten des Datenmodells sind in drei Hauptbereiche eingeteilt: die Definition eines Proteins und dessen taxonomische Zuordnung (grau), die Definition einer Sequenze sowie deren zugehörigen Annotationen (hellgrau) und die Definition einer Proteinstruktur (weiß).

Protein Entitäten

Ein Protein kann biochemisch definiert werden als eine Polymerkette bestehend aus einer eindeutigen Abfolge von Aminosäuren, die eine räumliche Struktur ausbildet. Diese Raumstruktur ist eng mit der biologischen Funktion und Wirkungsweise des Proteins gekoppelt. Diese primären Sequenz und Strukturdaten sind jedoch nicht hinreichend um ein Protein und dessen biologische Funktion voll zu erfassen. Um die Funktionsbeziehungen von Proteinen innerhalb einer Proteinfamilie zu untersuchen ist auch dessen biologische Herkunft, also der Ursprungsorganismus, von Bedeutung. Des weitern ist eine hierarchische Einteilung der Proteine in Gruppen basierend auf deren Sequenzähnlichkeit unablässig. Um diese, der Sequenz und Struktur übergeordneten Eigenschaften der Proteine zu erfassen wurden folgende Entitäten im Datenmodell definiert:

Proteine wurden aufgrund ihrer Sequenzähnlichkeit in 3 Hierarchieebenen eingeteilt: Die höchste Ebene stellte die Superfamilie dar. Superfamilien wurden in der Tabelle **S_Fam** durch den Primarschlüssel *Superfamily_ID* und dem Familiennamen *Superfamily* beschrieben. Die zweite Ebene stellte die homologen Familien dar, die in der Tabelle **H_Fam** durch den Primärschlüssel *H_Family_ID* und dem Familiennamen *H_Family_Name* beschrieben wurden. Jede homologe Familie wurde über den Fremdschlüssel *Superfamily_ID* einer Superfamilie zugeordnet. In der Tabelle **H_Fam_Ref_Seq** konnte eine Referenzsequenz für jede homologe Familie definiert werden. Die dritte Ebene ist das Protein selbst. Für jedes Protein wurde ein Datensatz mit dem Primärschlüssel *Protein_ID* in der Tabelle **Proteins** angelegt. Die Zuordnung zu einer homologen Familie erfolgte über den Fremdschlüssel *H_Family_ID*. Neben der Bezeichnung des Proteins in *Protein_Name* können Informationen über bekannte Abkürzungen in *Abbreviation* und allgemeine Bemerkungen zum Protein in *Remark* abgelegt werden. Theoretisch sollte jedem Protein eine einzige Aminosäuresequenz zuordenbar sein. In der praktischen Anwendung ist dieser Ansatz jedoch nicht durchführbar. In den öffentlich zugänglichen Datenbanken finden sich für ein Protein mehrere Einträge die sich in der Aminosäuresequenz unterscheiden können. Dies liegt zum einen darin begründet, dass die Ableitung der Proteinsequenz durch die Translation des zu Grunde liegenden Gens erfolgt. Bei der Bestimmung der Gensequenz können vereinzelt Sequenzierungsfehler auftreten, die zu abweichenden Proteinsequenzen führen. Des Weiteren ist die exakte Zuordnung des Ursprungsorganismus nicht immer möglich womit Proteinsequenzen aus unterschiedlichen Stämmen zusammengefasst werden. Ein weiteres Problem kann die Ableitung der Proteinsequenz aus der Proteinstruktur darstellen, da in flexiblen Bereichen der Struktur Aminosäuren vereinzelt nur teilweise oder gar nicht bestimmt werden können. Um

diesen Sachverhalt Rechnung zu tragen können im Datenmodell der LED einem Proteineintrag mehrere Aminosäuresequenzen zugeordnet sein, wobei die Sequenzidentität über 98% betragen musste. Zur besseren Handhabung der Proteineinträge wurde deshalb in der Tabelle **Protein_Ref_Seq** für jeden Proteineintrag eine Referenzsequenz definiert. Der Ursprungsorganismus eines Proteins wurde über die Tabellen **Genera** und **Species** beschrieben. Falls für den Organismus eine Stammbezeichnung bekannt war, wurde diese der Speziesbezeichnung angehängt und in *Species* abgelegt. Jede Spezies wurde über den Fremdschlüssel *Genus_ID* einem Genus zugeordnet.

Sequenz Entitäten

Dieser Bereich enthält Tabellen die zur Verwaltung der Aminosäuresequenz der Proteine und den an die Sequenz gebundenen Eigenschaften nötig sind, um effektive Analysen auf Sequenzebene zu ermöglichen. Durch einen Dublettenvergleich wurde ein nicht redundanter Satz an Sequenzen erzeugt wobei nicht nur die Sequenzidentität als Kriterium herangezogen wurde, sondern auch die Sequenzlänge. Somit wurde für ein Sequenzfragment ein getrennter Datensatz erzeugt, auch wenn dieses 100% Sequenzidentität zu einem längeren Sequenzeintrag aufwies. Eine Ausnahme stellten jedoch Sequenzen dar, die aus der Proteinstruktur abgeleitet wurden. Enthielt die Struktur mehrere Monomere wurde für jedes Monomer ein getrennter Sequenzeintrag erzeugt. Für jede Sequenz wurde in der Tabelle **Sequences** ein Datensatz mit dem Primärschlüssel *Sequence_ID* angelegt. Über den Fremdschlüssel *Protein_ID* erfolgte die Zuordnung zu einem Proteineintrag. *Seq_Description* enthält eine Beschreibung der Proteineigenschaften und in *putative* wurde festgehalten, ob es sich um ein experimentell validiertes Protein handelte. In der Tabelle **DB_Entries** wurden für jede Sequenz die Bezeichner der Einträge aus den Quelldatenbanken abgelegt und über den Fremdschlüssel *Sequence_ID* mit der Sequenz verknüpft. In den Quelldatenbanken wie z.B. im Fall der SWISS-PROT[88] sind Einträge häufig über zwei Bezeichner gekennzeichnet. Diese wurden in *DB_Entry* und *AC* abgelegt. Die Quelldatenbanken, die in der Tabelle **DB** definiert wurden, wurden über den Fremdschlüssel *DB_ID* verknüpft. Für jede Aminosäure wurde in der Tabelle **Positions** ein Datensatz erzeugt und durch den Primärschlüssel *Positions_ID* eindeutig bezeichnet. Die Position der Aminosäure in der Sequenz wurde in *AS_Position* und das Einbuchstabensymbol der Aminosäure in *AS_1* abgelegt. Dies hatte den Vorteil, dass mit Aminosäuren verbundene Annotationen leicht im Datenmodell angelegt werden konnten. Annotationen wurden in der Tabelle **Seq_Annotations** gespeichert. Annotationen, die an eine einzelne Aminosäure gebunden waren, wie z.B. katalytisch aktive Aminosäuren oder

Seitenketten, die an der Bindung von Substraten oder Cofaktoren beteiligt waren, wurden über den Fremdschlüssel *Position_ID* mit der Aminosäure verbunden. Für Annotationen, die in Abhängigkeit zu einer zweiten Aminosäure standen, wie z.B. Disulfidbrücken, wurde die zweite Aminosäure durch den Fremdschlüssel *Position_ID2* spezifiziert. *Comment* enthält nähere Informationen zur Annotation und die Annotation selbst wurde über den Fremdschlüssel *Description_ID* spezifiziert. Die Tabelle **Descriptions** beschreibt die für die untersuchte Proteinfamilie wichtigen Annotationen und kann für die jeweilige Proteinfamilie angepasst werden. Die für die α/β-Hydrolasen verwendeten Annotationen sind in Tabelle 1 beschrieben. *Description* enthält eine allgemeine Bezeichnung der Annotation, *Abbreviation* eine Abkürzung. *Visible* definiert, ob eine Annotation in Multisequenz Alignments dargestellt werden sollte. In *Color* wurde eine Farbe definiert, die zum Hervorheben der Annotation in generierten Multisequenz Alignments verwendet wurde und *Aln_Abbreviation* enthält ein Einbuchstabensymbol, das als Bezeichner der Annotation im Multisequenz Alignment diente. *Segmentt* definiert eine Annotation, die sich über einen Sequenzbereich erstreckte. Diese Information wurde genutzt, um neben der gesamten Sequenz auch die Möglichkeit zu bieten, nur diese Bereiche zur Erzeugung von Multisequenz Alignments einzubeziehen. Mutationsinformationen wurden getrennt von anderen Annotationen verwaltet. Die Tabelle **Mutations** enthält in *To_AS* den Aminosäureaustausch für die über den Fremdschlüssel *Positions_ID* gekennzeichnete Aminosäure. Eine genaue Beschreibung der durch die Mutation bewirkten Effekte finden sich in der Tabelle **M_Descriptions**. Auswirkungen auf die Struktur (*Structure_Effect*), experimentelle Beobachtungen (*Exp_Observations*) und Literaturstellen (*References*) wurden in dieser Tabelle abgelegt. Da in der LED ein nicht redundanter Satz an Sequenzen gehalten wurde, wurden Annotationen für eine Sequenz, die aus verschiedenen Einträgen unterschiedlicher Datenbanken stammten, auf einen Sequenzeintrag in der LED übertragen. Da der Verlauf von Annotationsänderungen nicht gespeichert wurde, konnten die ursprünglich extrahierten Annotationen nur über das Auswerten der original Einträge aus den Quelldatenbanken zurückverfolgt werden.

Tabelle 1 Annotationen für die Familie der α/β-Hydrolasen mit den zugehörigen Abkürzungen, die vor dem Beginn der Datenintegration initialisiert werden müssen. Die Groß- und Kleinschreibung der Datensätze ist zu beachten.

Description	Abbreviation	Alignment
active site	AS	
disulfide bridge	SS	
putative glycosylation site	G	G
signal-peptide	SP	
ER retenion signal	ERS	
oxyanion hole	O	O
anchor residue	A	A
binding site of medium sized moiety of a secondary alcohol	M	M
scissile fatty acid binding site	F	F
hydrophobic binding site or tunnel	H	H
lid	L	
metal binding site	MB	
propeptide position	P	
binding site	BS	
residue for superimposition	FIT	
codehop motif 1 (EDCL)	M1	
codehop motif 2 (VRENI)	M2	
proton donator	PD	
halide cradle	HC	
GXGXS	GGS	
transmembrane region	T	
not a/b hydrolase domain	nAB	
cap domain	CAP	

Struktur Entitäten

Das Datenmodell der LED zur Speicherung von Proteinstrukturdaten wurde an das der Protein Data Bank[89] (PDB) angelehnt. Dies erlaubte die Handhabung von experimentell bestimmten Proteinstrukturen als auch von theoretischen Modellen, sowie das leichte Überführen der gespeicherten Strukturdaten in das PDB Format, einem der gängigsten Eingabeformate für Strukturanalysesysteme. Enthielt ein Proteinstruktureintrag mehrere Monomere wurde für jedes Monomer ein Datensatz in der Tabelle **Structures** erzeugt. Über den Fremdschlüssel *Sequence_ID* wurde jeder Struktureintrag mit der zugehörigen Aminosäuresequenz verknüpft. Dies stellte die einzige Ausnahme von dem sonst nicht redundanten Satz an Sequenz der LED dar. Dieser Ansatz war notwendig, um aus der Proteinstruktur abgeleitete Eigenschaften eindeutig auf die Sequenz abbilden zu können. Da Annotationen im Datenmodel der LED auf Aminosäureebene definiert wurden, konnte nur so gewährleistet werden, dass z.B. die Sekundärstruktur zweier unterschiedlicher Proteinstrukturkonformere bei identischer Sequenz später dem entsprechenden

Struktureintrag zugeordnet werden konnten. Jeder Strukturdatensatz trägt in *Completeness* die Information darüber, ob Raumkoordinaten ausschließlich für C_α-Atome, das Peptidrückgrad oder auch die Seitenkettenatome bestimmt wurden. In *Conformation* wurde abhängig von der Proteinfamilie definiert welches Konformer vorlag. Im Fall der α/β-Hydrolasen sind die Zustände geöffnet, geschlossen und halb-geöffnet definiert. Zur Beschreibung der Herkunft des Struktureintrages wurden zwei Tabellen erstellt, **PDB** und **Models**. Für experimentell bestimmte Proteinstrukturen, die über die PDB[89] zugänglich waren, wurden Informationen über mögliche gebundene Inhibitoren (*Inhibitor*), den original PDB Bezeichner (*PDB_Entry*), die Auflösung der Struktur (*Resolution*), das Jahr der Veröffentlichung (*Year*) und die Autoren (*Authors*) in der Tabelle **PDB** gespeichert. Auf der LED basierende Homologiemodelle, die dem Modell zugrunde liegende Proteinstruktur (*Template*), die Bewertung des Models über Prosa[90] und die Methode der Strukturmodellierung (*Creation_method*) wurden in der Tabelle **Models** gespeichert. In Anlehnung an das PDB Datenmodell wurden Daten für Proteinatome, Heteroatome, Raumkoordinaten und speziellen Verbindungen zwischen Atomen in den Tabellen **Atoms**, **Hetatms**, **Coords** und **Connectivities** abgelegt. Die für jeden Struktureintrag berechneten Sekundär-strukturinformationen wurden in der Tabelle **DSSP** gespeichert.

3.2 Die Data Warehouse Architektur

Der konzeptionelle Aufbau des Data Warehouse Systems DWARF (Data Warehouse for Analysing Protein Families) ist in Abbildung 5 dargestellt. Dieser kann in 3 Bereiche eingeteilt werden: (1) Die Integration der Daten aus den Quelldatenbanken wird durch die ETL-Schnittstelle (ETL: Extrahieren, Transformieren, Laden) erzielt. Der Datenintegration schließen sich (2) die Bereinigung und Anreicherung sowie (3) die Analyse der integrierten Daten und der Datenzugriff an.

3.3 Extrahieren, Transformieren und Laden (ETL) von Daten

Der eigentlichen Datenintegration geht eine Initialisierung zweier Datenbanktabellen mit Standardwerten voraus, die von der ETL-Schnittstelle zur Transformation des NCBI Datenmodells in das der LED benötigt wurden. Dies betraf die Tabellen **DB** und **Descriptions**. Während die Informationen über die Quelldatenbanken in **DB** für alle

Proteinfamilien identisch waren, konnten die Annotationsbeschreibungen in **Descriptions** für verschiedene Proteinfamilien variieren.

Der Datenbankinitialisierung schloss sich die Datenintegration an. Dieser Vorgang bestand aus mehreren Iterationen zum Extrahieren, Transformieren und Laden sowie der Bereinigung und Anreicherung der Daten.

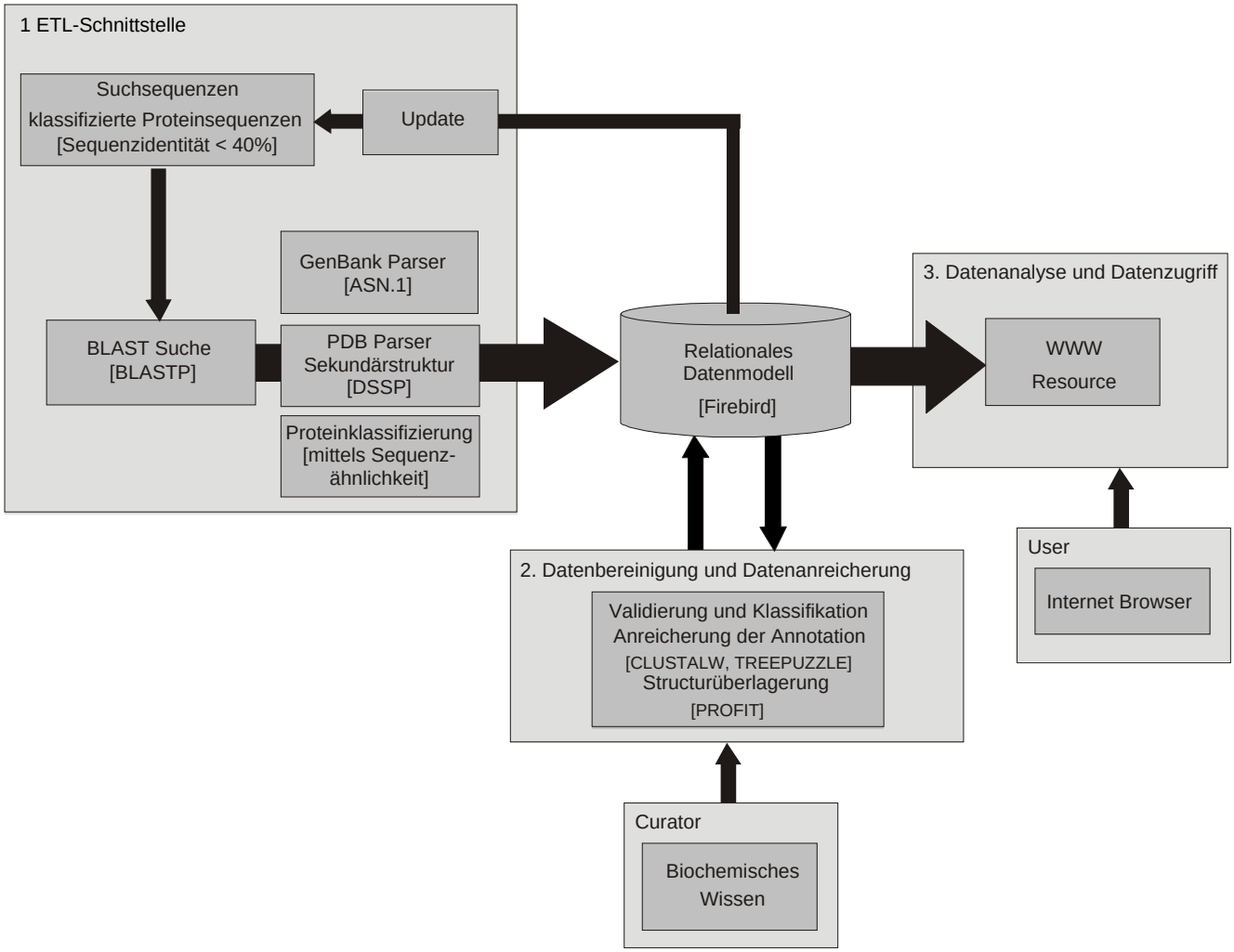


Abbildung 5 Die Architektur des Data Warehousesystems. Dieses ist in drei Bereiche eingeteilt: (1) Die ETL-Schnittstelle zum Extrahieren, Transformieren und Laden der Daten in die lokale Datenbank, (2) die Schnittstelle für den Datenbank *curator* zum Bereinigen und Anreichern der lokal gespeicherten Daten sowie (3) die Schnittstelle für den Benutzer zur Datensichtung und Analyse.

Extrahieren der Daten

Da die Identifikation von Proteineinträgen, die als Mitglieder der zu untersuchenden Proteinfamilie in Betracht kamen, über Homologiesuchen mittels BLAST[30] erfolgte, wurden für die erste Iteration der Datenextraktion vordefinierte Suchsequenzen benötigt. Diese ließen sich aus folgenden Quellen extrahieren: (1) Strukturklassifikationsdatenbanken wie SCOP[63] oder CATH[7] eigneten sich um Suchsequenzen innerhalb des bekannten Strukturraums der untersuchten Proteinfamilie zu identifizieren, (2) Sequenzclusterdatenbanken wie SYSTERS[91] oder (3) falls vorhanden aus auf die Proteinfamilie spezialisierte Datenbanken. Die so gewonnene Liste an Suchsequenzen wurde durch Stichwortsuchen in primären Sequenzdatenbanken wie GenBank[92] ergänzt. Um eine effiziente Datenintegration zu gewährleisten, wurden nur Suchsequenzen mit einer Sequenzähnlichkeit unter 40% gewählt, da ansonsten die BLAST Homologiesuchen eine erhöhte Redundanz in den Sequenztrefferlisten aufweisen. Auch wurden die Suchsequenzen aus einem breiten Spektrum an Organismen gewählt, um möglichst den gesamten bekannten Sequenzraum der Proteinfamilie abzudecken. Dies verringert die Anzahl der benötigten Iteration zur Vervollständigung der Datenbank.

Bevor die Homologiesuchen für die Suchsequenzen durchgeführt wurden, wurde für diese die entsprechenden Superfamilien und homologen Familien in der Datenbank angelegt und die Suchsequenzen in der Datenbank gespeichert. Sodann wurden für die einzelnen Vertreter der Superfamilien die Homologiesuchen in der nicht redundanten Proteinsequenzdatenbank des NCBI durchgeführt. Treffersequenzen werden automatisch existierenden homologen Familien zugeordnet. Hierzu wird für alle im Data Warehouse gespeicherten Proteinsequenzen eine lokale BLAST Datenbank erstellt, mit der die Treffersequenzen verglichen werden. Als Schwellenwert werden zwei E-Werte für die Zugehörigkeit zu einer Superfamilie oder homologen Familie definiert. Diese sind abhängig von der untersuchten Proteinfamilie. Für Treffersequenzen, die diese Schwellenwerte überschreiten wird eine neue Superfamilie bzw. homologe Familie angelegt.

Enthält eine Superfamilie mehrere homologe Familien, werden die Trefferlisten der Homologiesuchen nicht sequentiell, sondern parallel abgearbeitet. Dies bedeutet, dass die Verarbeitung der Treffersequenzen entsprechend dem Rang der Suchergebnisse für die einzelnen homologen Familien erfolgt. Diese Vorgehensweise verhindert das unausgewogene Anwachsen einzelner homologer Familien.

Für die Treffersequenzen werden die entsprechenden GenBank Einträge im ASN.1 Format des NCBI Datenmodells für die Datenextraktion heruntergeladen. Ein für die ASN.1 Notation

entwickelter Parser extrahiert aus diesen Einträgen Information über den Organismus, die Quelldatenbank, Annotationen, Mutationen und der Sequenz und transformiert diese in das Datenmodell der LED. Für Struktureinträge werden jedoch keine Informationen aus dem GenBank Eintrag extrahiert. Zwar wird der Eintrag klassifiziert, es wird aber nur ein Sequenzeintrag in **Sequences** erzeugt ohne die Aminosäuren in **Positions** abzulegen. Dies hat den Hintergrund, dass die in Genbank vorhandenen Einträge zu Proteinstrukturen die komplette Proteinsequenz enthalten, wogegen die original Struktureneinträge der PDB durch nicht aufgelöste Aminosäuren unvollständig sein können. Einträge für Strukturmonomere werden deshalb nach der Datenintegration der GenBank Einträge aus der ExPDB[93] bezogen und in die Datenbank integriert. Die Sekundärstruktur Informationen werden für jeden Strukturdatensatz mit DSSP[94] berechnet und ebenfalls in der Datenbank gespeichert.
Die hier beschriebene Methode wurde mit 3 Perl Skripten als automatisierter Prozess implementiert.

3.3.1 Datenbereinigung und Datenanreicherung

Jeder Iteration der Datenintegration schließt sich die Bereinigung und Anreicherung der Daten an. Hierfür wurde ein Webinterface entwickelt, das dem *curator* der Datenbank gestattet, die Daten zu sichten und zu manipulieren. Die Datenbereinigung beginnt mit der Überprüfung der Superfamilien und homologen Familien auf Konsistenz der zugewiesenen Proteinsequenzen. Multisequenz Alignments wurden mit CLUSTALW[95] für Superfamilien und homologe Familien erzeugt und hoch konservierte Sequenzmotive, sogenannte Familiendeskriptoren, wurden genutzt um Proteinsequenzen als Familienmitglieder zu identifizieren. Die Familiendeskriptoren werden in Datenbanken wie PROSITE [18] definiert oder können aus der Analyse der Daten des Data Warehouse eigenständig abgeleitet werden. Auf den Multisequenz Alignments basierend wurden mit Hilfe der Neighbour-Joining Clusterung, wie sie in CLUSTALW implementiert ist, Phylogenetische Bäume erstellt. Diese dienten zur Validierung der Klassifikation. Inkorrekt zugeordnete Proteinsequenzen konnten so identifiziert werden und wurden umklassifiziert. Da die Neighbour-Joining Clusterung nur eine distanzbasierte Methode zur phylogenetischen Analyse darstellt, wurde die endgültige Klassifikation anhand der Maximum Likelihood Methode überprüft, wie sie in TREE-PUZZLE[96] implementiert ist.
Die auf diesen konsistenten Familien basierenden Multisequenz Alignments wurden genutzt um die extrahierten Annotationen zu überprüfen und gegebenenfalls zu korrigieren. Literaturdaten und Ergebnisse aus eigenen Experimenten wurden zur Datenanreicherung genutzt. Das Markieren der Annotationen in den Multisequenz Alignments kann so zur

Qualitätsprüfung der Annotationen und der Alignments dienen, da strukturell und funktionell relevante Aminosäuren im Alignment konserviert sein sollten.

Um die Analyse der in der Datenbank gespeicherten Strukturdatensätze zur erleichtern wurden strukturell konservierte Bereiche in den Multisequenz Alignments identifiziert und manuell annotiert. Dieser Datenanreicherung schließt sich die automatisierte Überlagerung der Struktureinträge an. Die Überlagerung der Proteinstrukturen erfolgte nach dem McLachlan Algorithmus,[97] wie er in PROFIT (Martin, A.C.R., http://www.bioinf.org.uk/software/profit) implementiert ist.

3.3.2 Datenanalyse und Zugriff auf die Datenbank

Das Datenbankinterface für den Zugriff und die Analyse der Proteinfamilien ist in einen Bereich für Datenbankanwender und einen Bereich für den Datenbank *curator* aufgeteilt. Dem Datenbankanwender wird der Zugriff auf das Data Warehouse über eine Website ermöglicht, die die aufbereiteten Daten der Proteinfamilie in Form von statischen HTML Seiten bereitstellt. Der Zugang zu den einzelnen Familien wird über eine Übersichtsseite ermöglicht. Superfamilien und die zugehörigen homologen Familien werden tabellarisch dargestellt. Für jede Superfamilie sind die Anzahl der Protein-, Sequenz- und Struktureinträge, Links zu Multisequenz Alignments und phylogenetischen Bäumen sowie den untergeordneten homologen Familien erhältlich. Die Seiten der homologen Familien listen die zugehörigen Proteineinträge sortiert nach dem Organismus. Für jeden Proteineintrag sind Informationen über den Ursprungsorganismus, den Proteinnamen, eine Beschreibung, Links zu den zugehörigen Sequenzeinträgen in GenBank und falls vorhanden Links zu den überlagerten Struktureinträgen im PDB Format zugänglich. Des Weiteren werden Multisequenz Alignments und phylogenetische Bäume bereitgestellt. Nur die Referenzsequenzen der Proteineinträge gehen in die Multisequenz Alignments ein und werden mit den Zugangscodes der Quelldatenbank bezeichnet. Die im Data Warehouse gespeicherten Annotationen werden in den Proteinsequenzen des Alignments farbig markiert und die Selektion der markierten Aminosäuren gibt Informationen über alle mit dieser Position verbundenen Annotationen zurück. Für Proteinstrukturen wird die Sekundärstrukturinformation im Alignment dargestellt. Die für Superfamilien und homologe Familien erzeugten phylogenetischen Bäume wurden mit TREE-PUZZLE[96] erstellt und die Blätter mit den Zugangscodes der Quelldatenbank bezeichnet und mit diesen verknüpft. Die statistische Signifikanz ist für jede Verzweigung durch eine Bootstrap Analyse gegeben.

Zusätzlich hat der Datenbankanwender die Möglichkeit über eine BLAST Suche eine Sequenz mit allen in der Datenbank gespeicherten Sequenzen zu vergleichen. Die

Homologiesuche liefert eine Liste der ähnlichsten Sequenzen mit dem Verweis auf die zugehörigen Superfamilien und homologen Familien zurück.

Dem Datenbank *curator* steht ein zusätzliches, webbasiertes Interface für die Datenbereinigung und Datenanreicherung zur Verfügung. Dem *curator* stehen 3 Bereiche zur Auswahl (Abbildung 6): (1) der Menüpunkt Families ermöglicht den Zugang zu den Superfamilien und homologen Familien der Proteinfamilie, (2) der Bereich BLAST erlaubt die Verwaltung der lokalen BLAST Datenbank sowie Homologiesuchen gegen die Datenbank und (3) der Bereich Structures, der einen gezielten Zugang zu Struktureinträgen darstellt.

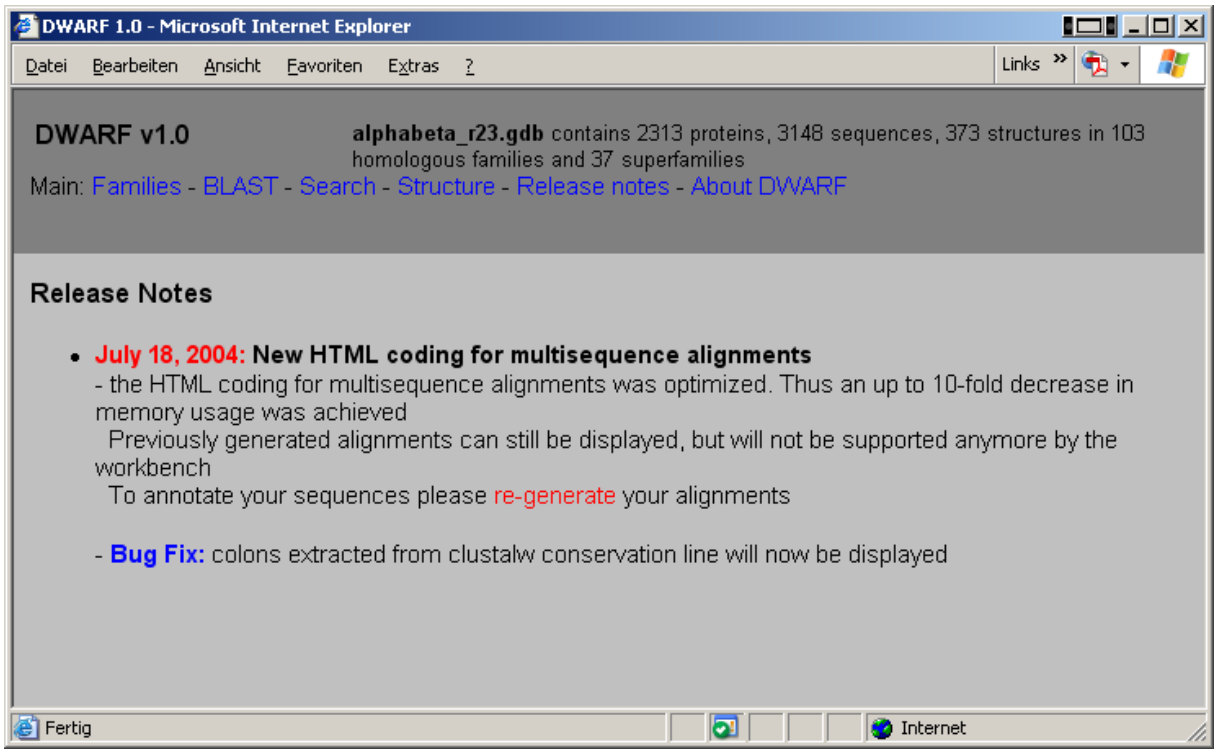


Abbildung 6: Startseite des Web Interface für den Datenbank *curator*. Es bestehen Menüpunkte zur Sichtung der Superfamilien und homologen Familien (Families), zur lokalen Sequenzähnlichkeitssuche (BLAST) und zur Sichtung der Strukturinformationen (Structure).

Families: Dieser Bereich stellt den zentralen Zugang zum Data Warehouse dar. Sämtliche Werkzeuge zur Datenbereinigung und zur Datenanreicherung auf Sequenzebene sind hier zugänglich. Zur Selektion der zu bearbeitenden Proteinsequenzen wird der *curator* durch die hierarchische Organisation der Proteineinträge geführt. Den Einstiegspunkt stellt eine Liste aller homologen Familien dar, die nach dem Namen der zugehöriger Superfamilie oder aber dem Namen der homologen Familie sortiert werden kann (Abbildung 7).

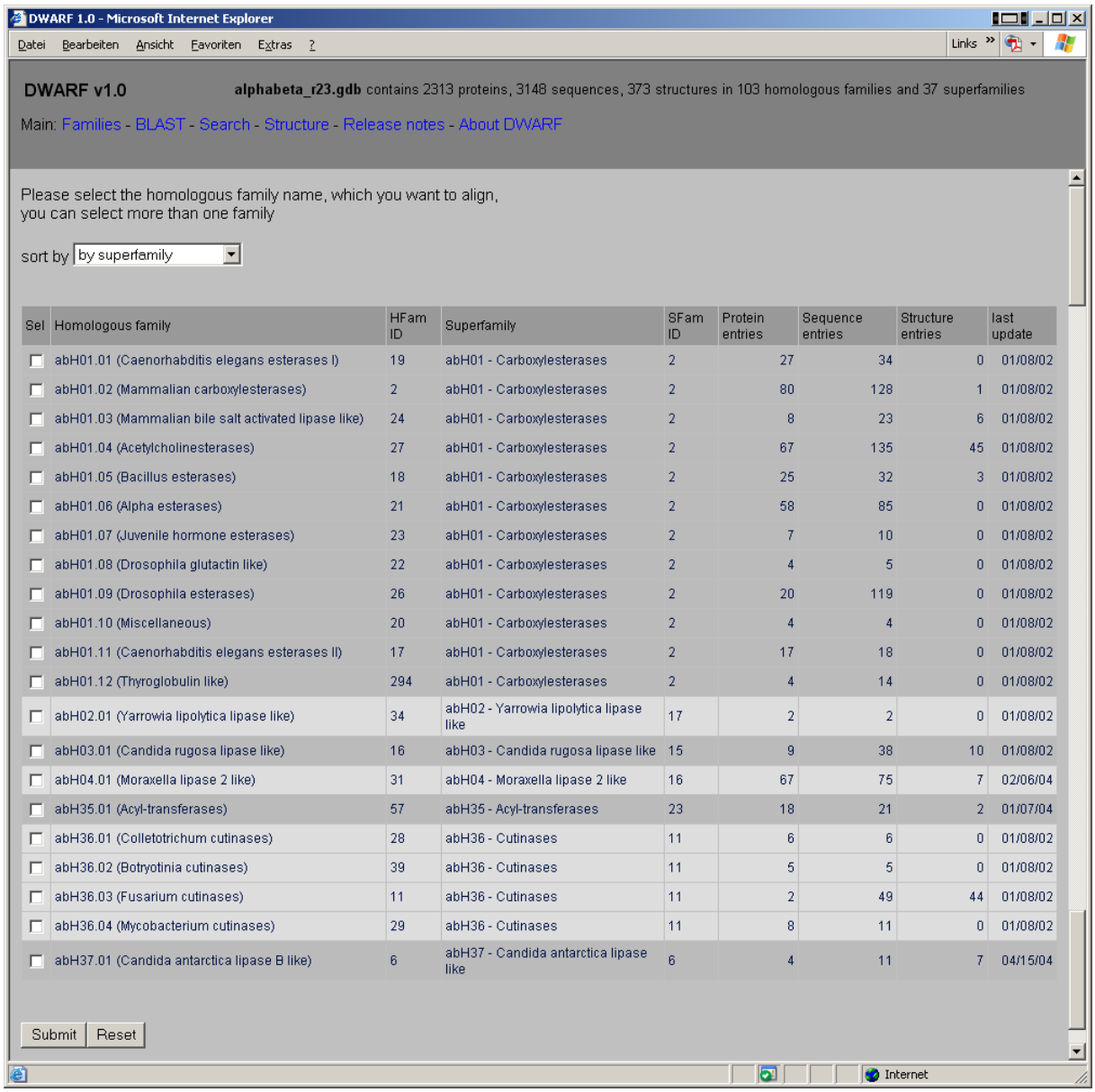


Abbildung 7: Die Übersicht der homologen Familien stellt den zentralen Zugang zu allen Werkzeugen zur Datenbereinigung und Datenanreicherung auf der Sequenzebene dar. Diese Ansicht repräsentiert die Namen der homologen Familien und Superfamilien, deren datenbankinternen Bezeichner sowie die Anzahl der Protein-, Sequenz- und Struktureinträge.

Nach der Selektion einer oder mehrerer homologen Familien wird eine Übersicht aller diesen Familien zugehörigen Proteineinträge erzeugt (Abbildung 8). Diese Übersicht hat die Funktion, die Proteineinträge zu sichten und zu editieren, Proteinsequenzen für ausgewählte Proteine im FASTA Format zu speichern und annotierte Multisequenz Alignments zu erzeugen, die zur Bereinigung und Anreicherung sowie zur Überprüfung der Klassifikation dienen. Die Proteineinträge werden nach der homologen Familie und dem Ursprungsorganismus sortiert dargestellt. Informationen über den Proteinnamen, den

Proteinbezeichner, Ursprungsorganismus und den untergeordneten Sequenzeinträgen sind zugänglich. Der Proteinname kann durch Selektion geändert werden. Zum Umklassifizieren eines Proteineintrags in eine andere homologe Familie ist über den Proteinbezeichner ein Werkzeug verknüpft, das die Auswahl der neuen homologen Familie, der das Protein angehören soll, ermöglicht. Für die Sequenzeinträge sind Informationen über den Sequenzbezeichner, den Namen der homologen Familie und deren Bezeichner, den Quelldatenbanken und den entsprechenden GenBank Einträgen, aus denen Informationen extrahiert wurden, der Klassifikation als putatives Protein sowie einer Beschreibung des Sequenzeintrages zugänglich. Mit den Sequenzbezeichnern ist ein Werkzeug zum Bearbeiten des Sequenzeintrags verknüpft (Abbildung 9).

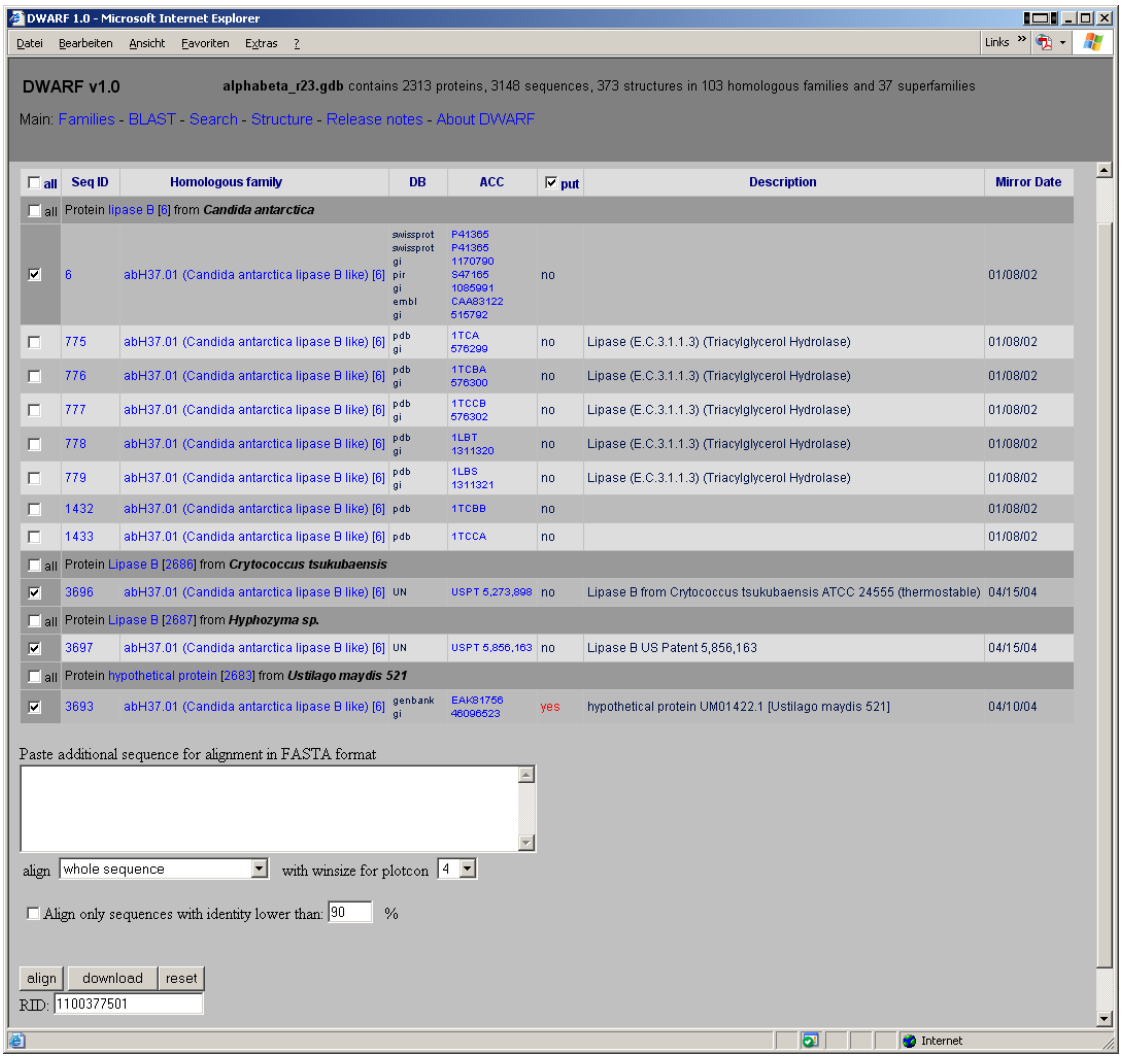


Abbildung 8: Übersicht der Proteineinträge einer homologen Familie am Beispiel der Familie abH37.1. Für jeden Proteineintrag wird der Name, der datenbankinterne Bezeichner und der Ursprungsorganismus angezeigt (dunkelgrau unterlegte Zeilen). Für die zugehörigen Sequenzeinträge werden die datenbankinternen Bezeichner, die homologe Familie, die Quelldatenbanken, die zugehörigen GenBank Einträge und eine Beschreibung angezeigt. Für jeden Proteineintrag ist eine vorselektierte Referenzsequenz definiert (dunkelgrau markierte Sequenzzeile). Für die markierten Sequenzeinträge kann ein Multisequenz Alignment erzeugt werden.

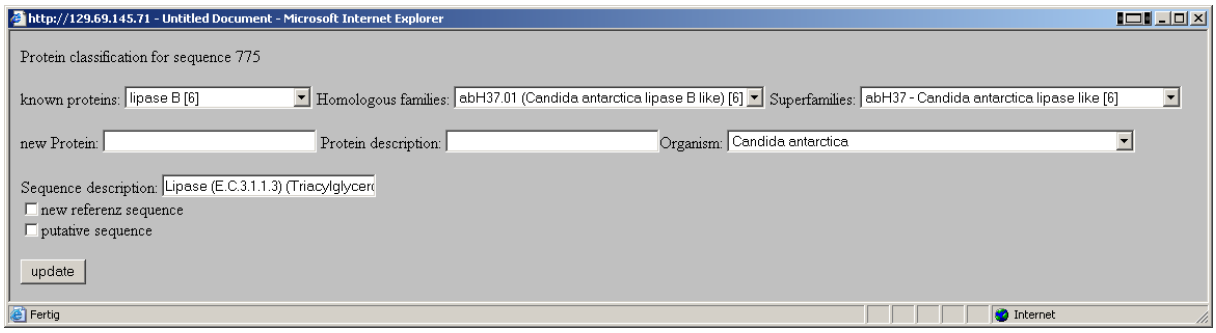


Abbildung 9: Werkzeug zur Bearbeitung eines Sequenzeintrags. Ermöglicht die Editierung der Attribute Proteinzugehörigkeit, Referenzsequenz und putative Sequenz.

Mit diesem Werkzeug kann der Sequenzeintrag einem anderen Proteineintrag zugeordnet werden, die Sequenzbeschreibung kann editiert werden und die Proteinsequenz kann als putativ und Referenzsequenz definiert werden.

Für jeden Proteineintrag wird ein Sequenzeintrag als Referenz definiert. Dieser ist in der Proteinübersicht vorselektiert. Für die Erstellung der annotierten Multisequenz Alignments können zusätzlich zu den ausgewählten Sequenzeinträgen weitere Proteinsequenzen im FASTA Format eingefügt werden. Sind in der Tabelle **Descriptions** des Datenmodells über das Attribut *Segmentt* Sequenzbereiche definiert, kann die Erstellung des Multisequenz Alignments auf diese Bereiche beschränkt werden. Dies ermöglicht z.B. im Fall von Multidomainproteinen nur die Domäne für Untersuchungen heranzuziehen, die der untersuchten Proteinfamilie angehört. Für jedes Multisequenz Alignment wird mit dem Algorithmus, wie er in PLOTCON des EMBOSS Programmpakets[98] implementiert ist ein Ähnlichkeitswert für die Konservierung der Aminosäurepositionen berechnet (Abbildung 10).

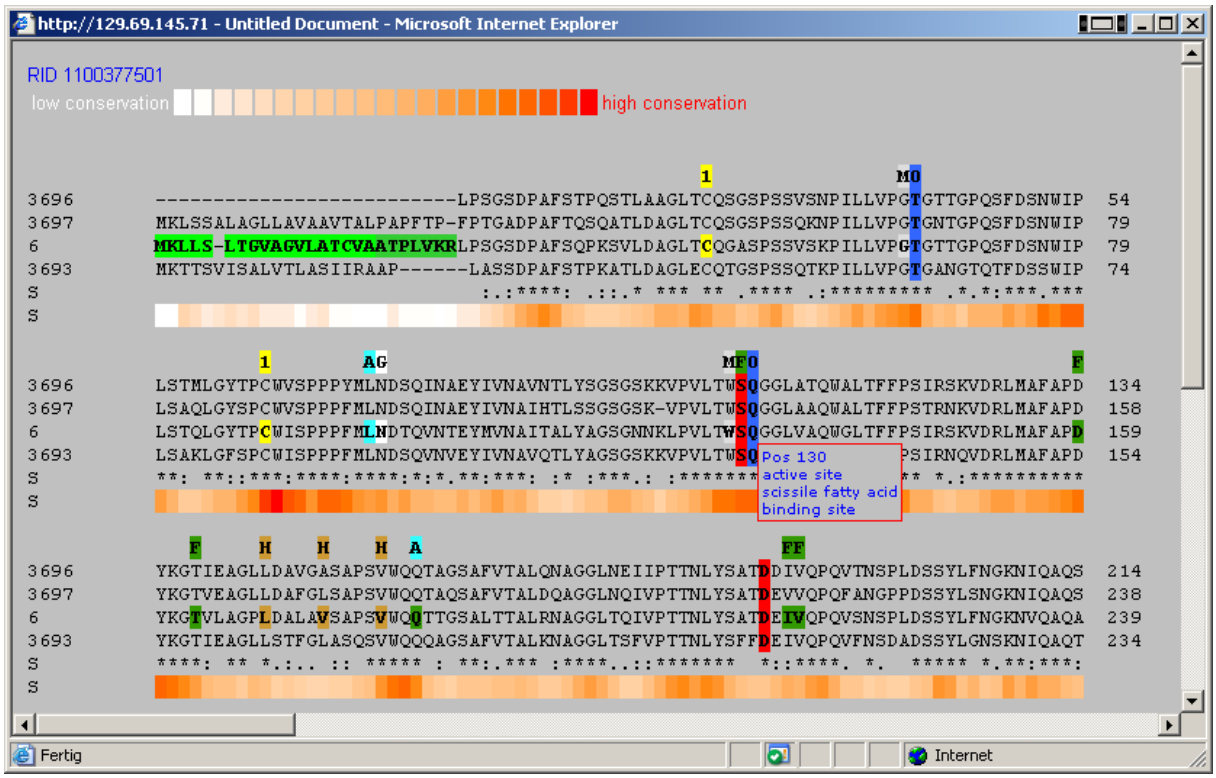


Abbildung 10: Darstellung eines über DWARF erstellten Multisequenz Alignment am Beispiel der Referenzsequenzen der Familie abH37.1. Aminosäurespezifische Informationen, die in der Datenbank gespeichert sind werden auf das Multisequenz Alignment aufgeblendet (farbig markierte Aminosäuren). Das Ergebnis einer Konservierungsanalyse wird als Farbspektrum (weiß: geringe Konservierung, rot: hohe Konservierung) jeweils in der letzten Zeile eines Blocks dargestellt.

Neben dem Multisequenz Alignment wird auch ein Neighbour joining Baum dargestellt, der zur Überprüfung der Klassifikation dient. Die Blätter des Baums sind mit den Sequenzbezeichnern versehen, die mit dem Werkzeug zur Umklassifikation des Proteineintrags verknüpft sind. Zur Bereinigung und Anreicherung der Sequenzannotationen steht ein weiteres Werkzeug bereit, das automatisch mit dem Multisequenz Alignment geöffnet wird. Nach der Selektion des zu bearbeitenden Sequenzeintrags öffnet sich das Annotationswerkzeug (Abbildung 11). Aminosäuren in der Proteinsequenz können ausgewählt werden um Annotationen zu ergänzen oder zu korrigieren. Zum effizienten Annreichern der Annotation stehen auch Funktionen bereit, um einzelne oder alle Annotationen einer Sequenz auf alle im Multisequenz Alignment vorhandenen Proteinsequenzen zu übertragen.

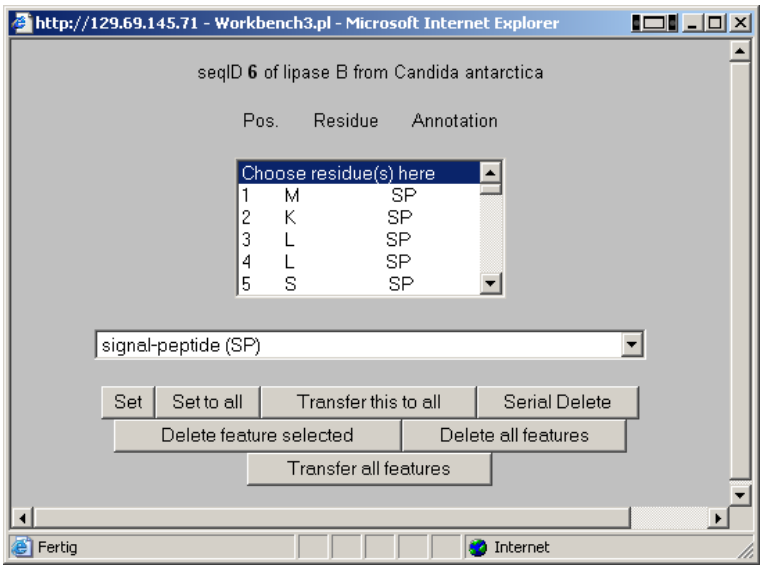


Abbildung 11: Werkzeug zur Bereinigung und Anreicherung von Sequnezannotationen. Aminosäuren des ausgewählten Sequenzeintrags können markiert und für diese Annotationen hinzugefügt oder entfernt werden. Diese Informationen können gleichzeitig auf im Multisequenz Alignment äquivalente Positionen übertragen werden.

BLAST: Dieser Bereich erlaubt dem Datenbank c*urator* das Erstellen der lokalen BLAST Datenbank sowie die Suche in dieser (Abbildung 12). Die lokale BLAST Suche kann bei der Überprüfung der Klassifikation von entscheidendem Vorteil sein. Für vom Parser automatisch erzeugte Familien kann so überprüft werden, ob eine Zugehörigkeit zu einer anderen Superfamilie möglich ist. Über den Menüpunkt „do BLAST“ erhält der *curator* die Möglichkeit eine Sequenz gegen das Data Warehouse zu blasten. Da die BLAST Datenbank nicht direkt im Datenmodell implementiert ist, ist ein regelmäßiges Erstellen der BLAST Datenbank über den Menüpunkt „build BLAST Database“ notwendig, um ein konsistentes Ergebnis zu gewähren.

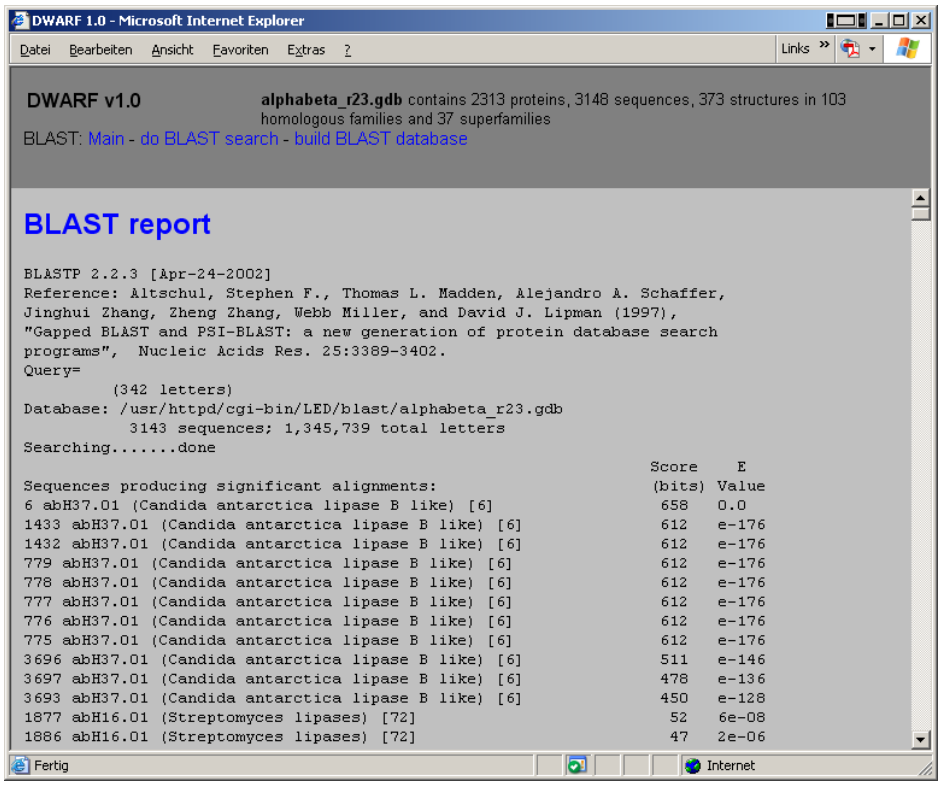


Abbildung 12: Ergebnis einer Sequenzähnlichkeitssuche in der LED mit dem in BLAST implementierten Algorithmus.

Structure: Dieser Bereich erlaubt den direkten Zugang zu den Proteinstruktureinträgen um diese gezielt zu analysieren und zu bearbeiten. Eine erste Auswahl erfolgt über die Selektion von homologen Familien oder Proteineinträgen, für die im Data Warehouse Strukturdaten abgelegt sind. Für die Selektion werden die gespeicherten Strukturdateneinträge zurückgeliefert. Informationen über Auflösung der Struktur, Autor und Veröffentlichungsjahr sowie zugehörige Familie, Organismus und Proteineintrag sind zugänglich. Für die aus dieser Liste ausgewählten Einträge, kann ein Multisequenz Alignment erstellt werden, das zur Identifikation von strukturell erhaltenen Bereichen genutzt werden kann. Diese können mit Hilfe des Annotationswerkzeuges definiert werden. Speziell das Anlegen des Aminosäuresatzes, der zur Überlagerung der Proteinstrukturen genutzt wird, ist hier möglich. Für die Überlagerung selbst muss eine Referenzstruktur selektiert werden, die als Vorlage für die Überlagerung genutzt wird.

3.4 Die α/β-Hydrolase Fold Datenbank

Das hier vorgestellte Data Warehouse System wurde genutzt um die *Lipase Engineering Database* (LED) aufzusetzen. Ursprünglich als Werkzeug zur Analyse der Enzymfamilie der Lipasen gedacht, ist dieses auf die gesamte Familie der α/β-Hydrolasen erweitert worden.

3.4.1 Der Sequenzraum der α/β-Hydrolasen

Die LED enthielt im Release 2.3 3148 Sequenzeinträge für 2313 Proteineinträge wobei 35% der Proteine als putative definiert waren. Für 96 Proteineinträge waren 261 Struktureinträge in der PDB zugänglich mit einer Gesamtanzahl von 373 Strukturdatensätzen. Die α/β-Hydrolasen wurden in 37 Superfamilien und 103 homologe Familien eingeteilt. Hierbei unterschieden sich die Superfamilien in der Anzahl ihrer Proteine deutlich (Abbildung 13). Die Superfamilie abH8 war mit 364 Proteinen die größte Familie und umfasste mehr als 16% aller Proteineinträge. Die drei größten Superfamilien abH8, abH1 und abH34 enthielten insgesamt 897 Proteine und stellten 39% der gesamten Datenbank dar.

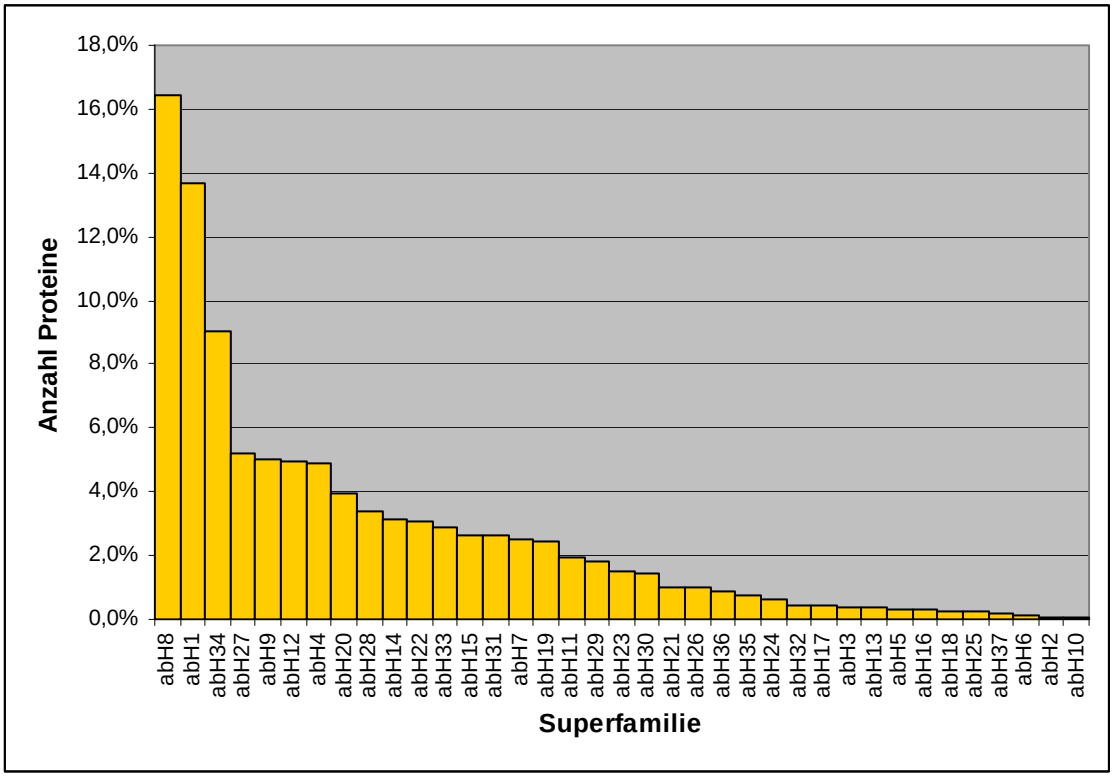


Abbildung 13: Verteilung der Anzahl der Proteineinträge über die Superfamilien der LED. Für die insgesamt 37 Superfamilien enthält die Superfamilie abH8 über 16% aller Proteineinträge. Die drei größten Superfamilie abH8, abH1 und abH34 enthalten mehr als 40% aller Proteineinträge.

Die Länge der Referenzsequenzen für diese Proteineinträge variierte deutlich. Betrachtete man die gesamten Sequenzlängen einschließlich Multidomainproteine so reichte diese von 33 bis 2769 Aminosäuren (Abbildung 14). Hierbei besaßen 62 Proteineinträge eine Sequenzlänge kürzer als 150 Aminosäuren und 310 Proteineinträgen mit einer Sequenzlänge größer als 650 Aminosäuren. Betrachtete man jedoch nur die Länge der korrespondierenden α/β-Hydrolase Domäne (Abbildung 15) zeigte sich, dass nur noch 54 Proteineinträge eine Sequenzlänge größer als 650 Aminosäuren aufwiesen. Somit lagen 3% der Proteineinträge oberhalb des Bereiches für die Sequenzlänge des α/β-Hydrolase Folds, der durch die Lipase A aus *Bacillus subtilis* mit 181 Aminosäuren und der Kokainesterase verwandten Hydrolase aus *Xanthomonas citri* mit 615 Aminosäuren Länge mit bekannter Proteinstruktur beschrieben wird. Diese 3% umfassten neben den Neuroligin- und Gliotactineinträgen der Superfamilie abH1 (8 Proteine) überwiegend hypothetische Proteine aus *C. elegans* (13 Proteine) und putative hormonsensitive Lipasen (6 Proteine).

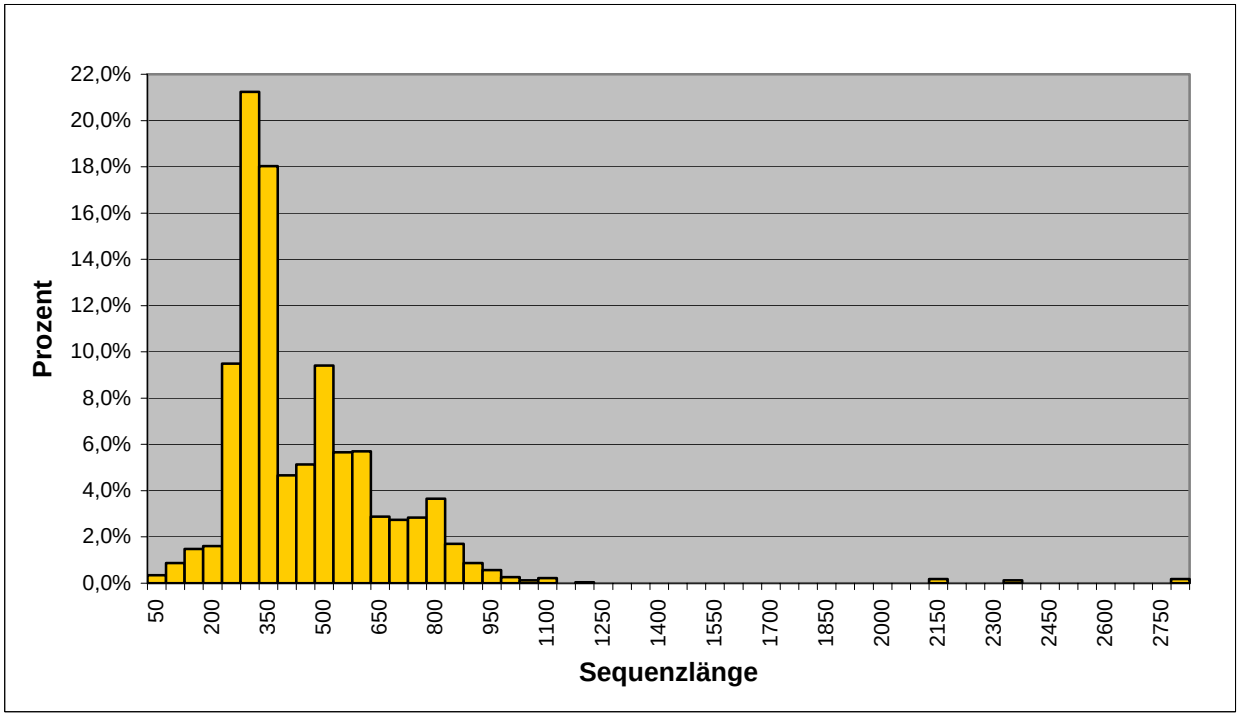


Abbildung 14: Histogramm der Sequenzlänge für die Proteinreferenzsequenzen der α/β-Hydrolasen einschließlich Multidomänenproteine.

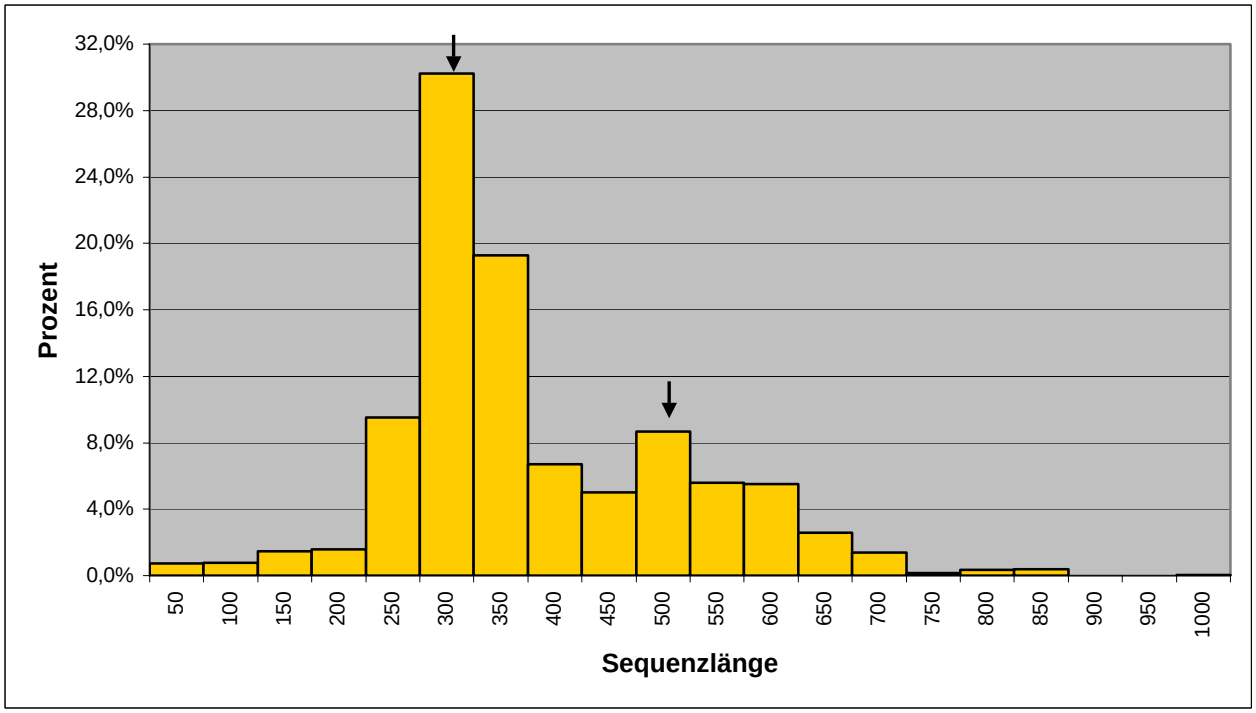


Abbildung 15: Histogramm der Sequenzlänge der α/β-Hydrolase Domäne der Referenzsequenzen. Die Modalwerte bei einer Länge von 300 und 500 Aminosäuren sind mit einem Pfeil markiert.

Die bimodale Verteilung der Sequenzlängen der α/β-Hydrolase Domänen wies einen Modalwert bei einer Länge von 300 Aminosäuren und einen weiteren Modalwert bei 500 Aminosäuren auf (Abbildung 15). Die Untersuchung der mittleren Sequenzlängen innerhalb der Superfamilien (Abbildung 16) zeigte, dass während für die Sequenzlänge um 300 Aminosäuren mehrere Familien beteiligt waren, für die Sequenzlänge um 500 Aminosäuren nur die Superfamilien abH1 und abH30 wesentlich beitrugen.

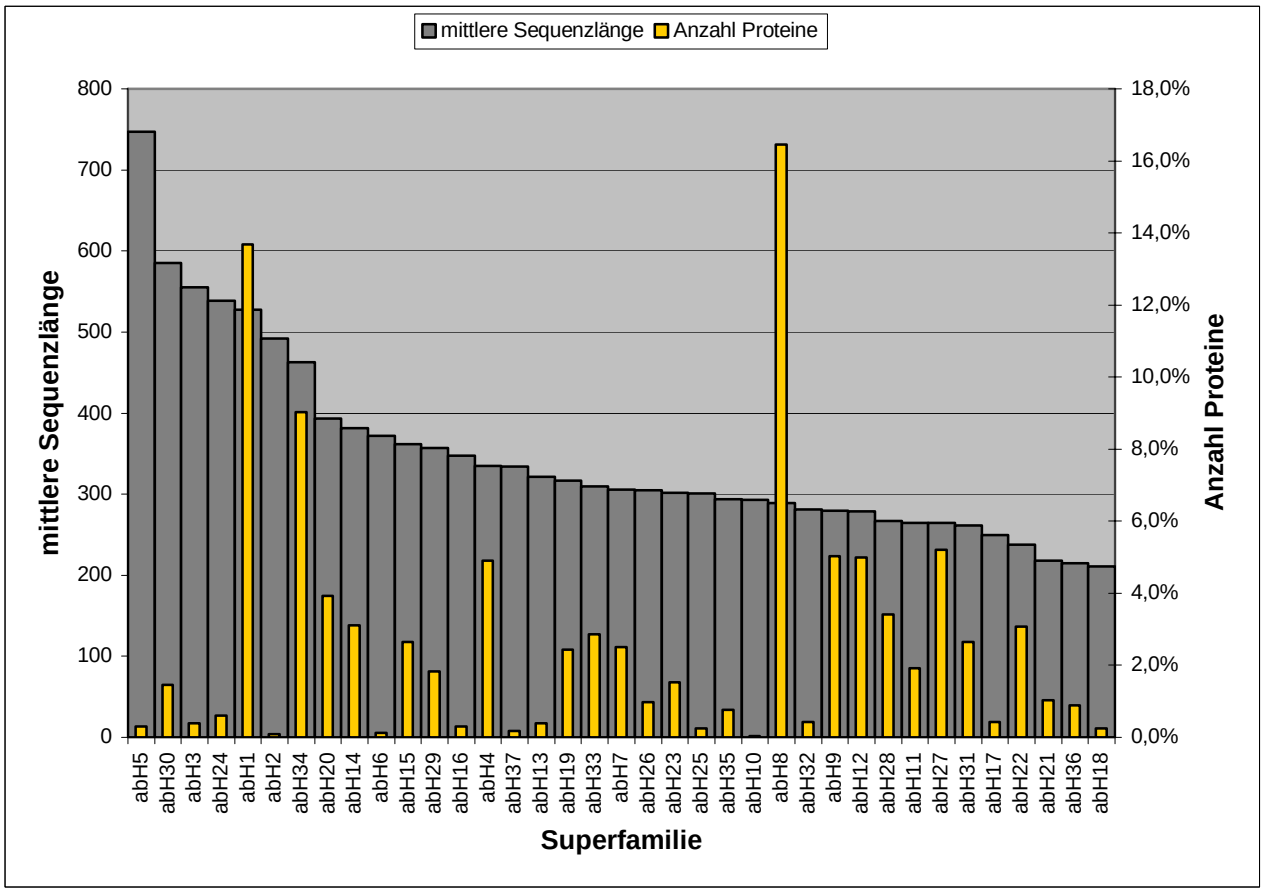


Abbildung 16: Vergleich der mittleren Sequenzlänge pro Superfamilie (grau) zur Anzahl der Proteine pro Superfamilie (gelb).

3.4.2 Drei α/β-Hydrolase Klassen

Struktur- und Sequenzanalysen ermöglichten die Einteilung der Proteineinträge in drei Klassen, die GX-, GGGX- bzw. Y-Klasse. Diese Klassenzuordnung basierte auf der Architektur des *oxyanion holes* und wird in Absatz 3.5.6 beschrieben. Die GX Klasse bestand aus 27 Superfamilien und 79 homologen Familien, die 1577 Proteineinträge mit 2015 Proteinsequenzen und 258 Strukturdatensätzen umfassten. Diese Klasse beinhaltete überwiegend bakterielle Lipasen und Lipasen aus Pilzen, eukaryotische Lipasen (hepatische, lipoprotein, pankreatisch, gastrische und lysosomale saure Lipasen), Cutinasen, Phospholipasen, hämfreie Peroxidasen, Acyltransferasen, Epoxidhydrolasen, Deacetylasen, Hydroxynirtrillyasen und Thioesterasen.

Die GGGX Klasse bestand aus 6 Superfamilien und 20 homologen Familien, die 457 Proteineinträge mit 793 Proteinsequenzen und 74 Strukturdatensätzen, umfassten. Diese

Klasse beinhaltete bakterielle Esterasen, α-Esterasen, eukaryotische Carboxylesterasen, Bile-salt aktivierte Lipasen, Juvenile Hormon Esterasen, hormon-sensitive Lipasen, Acetylcholinesterasen, Brefeldin A Esterasen und Thioesterasen.
Die Y Klasse besteht aus 4 Superfamilien und 4 homologen Familien, die 279 Porteineinträge mit 340 Proteinsequenzen und 41 Strukturdatensätzen, umfassten. Diese Klasse beinhaltete überwiegend Dipeptidylpeptidasen, Prolylendopeptidasen und Kokainesterasen.

3.5 Verwandtschaftsverhältnisse innerhalb der Lipasen, HMMs

Der Sequenzraum der α/β-Hydrolasen ist hoch divers und es ist nicht offensichtlich, ob α/β-Hydrolasen eine homologe oder analoge Foldfamilie darstellen. So lassen sich z.B. 54% der Aminosäuren der Lipase B aus *Candida antarctica* mit bekannter Proteinstruktur mit der Struktur der Lipase aus *Bacillus subtilis* mit einem RMSD von 1,23 Å überlagern, die Sequenzidentität beträgt für das Struktur Alignment jedoch nur 10%. Um die Verwandtschaft zwischen den Superfamilien zu untersuchen wurden HMM Profile für jede homologe Familie erstellt und der Verwandtschaftsgrad über Profil HMM-HMM Vergleiche ermittelt. Auf den aus diesen Vergleichen gewonnen Wahrscheinlichkeitswerten P konnte ein distanzbasiertes Netzwerk von Verwandtschaftsbeziehungen innerhalb der α/β-Hydrolasen gewonnen und ein hierarchischer Baum über die distanzbasierte UPGMA Methode abgeleitet werden (Abbildung 17).
Auf diesen Baum basierend wurde eine systematische Nomenklatur für die Familie der α/β-Hydrolasen erstellt. Jeder Familie wurde ein systematischer Name abHn.m zugewiesen, wobei abH (Akronym für α/β-Hydrolase Fold) von der Superfamilie n gefolgt wird, die von der homologen Familie m durch einen Punkt getrennt ist.
Die homologe Familie abH4.1 stellt jedoch eine Ausnahme dar und wurde nicht anhand der Baumtopologie eingeordnet. Diese Familie, die Proteine ähnlich zur Moraxella Lipase 2 enthält, ist aufgrund der Profil HMM-HMM Vergleiche zu den Superfamilien abH1 bis abH3 verwandt. Die höhere Konservierung der katalytischen Triade in Multisequenz Alignments mit Vertretern der Superfamilie abH4 im Vergleich zu abH2 führte jedoch zu diese Umklassifikation.
Der hierarchische Baum besteht aus sechs Hauptästen: (1) Familie abH1 bis abH6, die die GGGX Klasse repräsentierten (siehe Absatz 3.5.6), (2) Familie abH7 bis abH31, die den Grossteil der Familien der GX Klasse sowie sämtliche Y Klassefamilien (abH27 bis abH30) umfassten, (3) Familie abH32 bis abH34 umfasste Xylanase Esterasen, Antigen 85 und

Lysosomal protective Proteine, (4) abH35 Vertreter der Acyl Transferasen, (5) abH36 stellt die Familie der Cutinasen dar, sowie (6) abH37, die die Lipase B aus *Candida antarctica* enthält.

Für die Familie der Haloacid Dehalogenasen existiert keine Proteinstruktur, womit eine eindeutige Zuordnung zu einer Foldfamilie schwierig war. Über BLAST konnte eine lokale Ähnlichkeit zur Familie abH8.3 identifiziert werden. Der Vergleich der HMM Profile ergab jedoch keine signifikanten Ähnlichkeiten zwischen dieser Familie und den α/β-Hydrolasen, weshalb eine Einordnung dieser Familie nicht erfolgte.

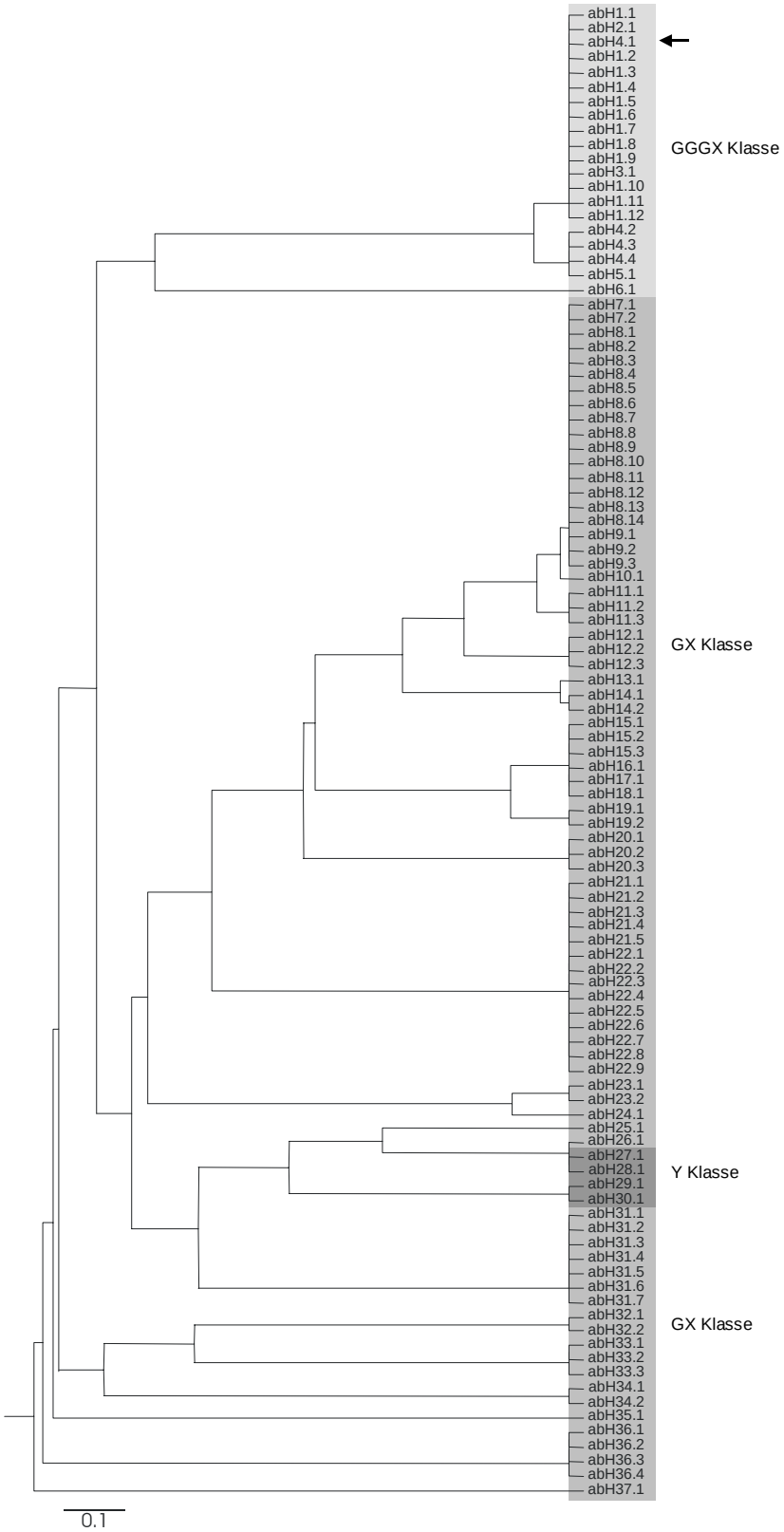


Abbildung 17: Verwandtschaftsbeziehung der homologen Familien abgeleitet aus Profil HMM-HMM Vergleichen. Anhand dieses hierarchischen Baums wurde die systematische Nomenklatur der α/β-Hydrolasen eingeführt. Die übergeordneten Klassen GGGX (hellgrau), GX (grau) und Y (dunkelgrau) leiten sich aus der Architektur des *oxyanion holes* ab. Die manuell umklassifizierte homologe Familie abH4.1 ist mit einem Pfeil markiert.

Tabelle 2: Liste aller in der LED (Release 2.3) abgelegten homologen Familien mit zugehöriger Superfamilie und Klasse. Sowohl die systematische Nomenklatur als auch eine beschreibende Bezeichnung sind angegeben.

Klasse	Superfamilie	Homologe Familie
GGGX	abH1 - Carboxylesterases	abH1.1 (Caenorhabditis elegans esterases I)
		abH1.2 (Mammalian carboxylesterases)
		abH1.3 (Mammalian bile salt activated lipase like)
		abH1.4 (Acetylcholinesterases)
		abH1.5 (Bacillus esterases)
		abH1.6 (Alpha esterases)
		abH1.7 (Juvenile hormone esterases)
		abH1.8 (Drosophila glutactin like)
		abH1.9 (Drosophila esterases)
		abH1.10 (Miscellaneous)
		abH1.11 (Caenorhabditis elegans esterases II)
		abH1.12 (Thyroglobulin like)
GGGX	abH2 - Yarrowia lipolytica lipase like	abH2.1 (Yarrowia lipolytica lipase like)
GGGX	abH3 - Candida rugosa lipase like	abH3.1 (Candida rugosa lipase like)
GGGX	abH4 - Moraxella lipase 2 like	abH4.1 (Moraxella lipase 2 like)
		abH4.2 (Acinetobacter esterases)
		abH4.3 (N-deacetylases)
		abH4.4 (Bacillus sphaericus lipase like)
GGGX	abH5 - Hormone sensitive lipases	abH5.1 (Hormone sensitive lipases)
GGGX	abH6 - Brefeldin A esterase like	abH6.1 (Brefeldin A esterase like)
GX	abH7 - Moraxella lipase 3 like	abH7.1 (Haemophilus influenzae lipase like)
		abH7.2 (Moraxella lipase 3 like)
GX	abH8 - Cytosolic Hydrolases	abH8.1 (soluble bacterial epoxide hydrolases I)
		abH8.2 (soluble bacterial epoxide hydrolases II)
		abH8.3 (soluble mammalian epoxide hydrolases)
		abH8.4 (soluble plant epoxide hydrolases)
		abH8.5 (soluble haloalkane dehalogenases)
		abH8.6 (miscellaneous)
		abH8.7 (soluble epoxide hydrolases (beta6))
		abH8.8 (soluble meta cleavage compound hydrolases II)
		abH8.9 (soluble non-heme peroxidases)
		abH8.10 (soluble haloalkane dehalogenases (beta6))
		abH8.11 (soluble meta cleavage compound hydrolases I)
		abH8.12 (Enol-lactone hydrolases)
		abH8.13 (soluble esterases / lipases / peptidases)
		abH8.14 (Ccg1/TafII250-interacting factor B like)
GX	abH9 - Microsomal Hydrolases	abH9.1 (microsomal epoxide hydrolases)
		abH9.2 (BioH protein like)
		abH9.3 (Proline iminopeptidases)
GX	abH10 - Uncultured crenarchaeote	abH10.1 (uncultured crenarchaeote)
GX	abH11 - Carboxylesterases	abH11.1 (Bacillus carboxylesterases)
		abH11.2 (Mycoplasma lipases)
		abH11.3 (Thermotoga maritima esterase like)
GX	abH12 - Hydroxynitrile lyases	abH12.1 (Hydroxynitrile lyases)
		abH12.2 (Arabidopsis thaliana esterases II)
		abH12.3 (Arabidopsis thaliana esterases I)
GX	abH13 - Bacterial esterases	abH13.1 (Bacterial esterase)
GX	abH14 - Gastric lipases	abH14.1 (Lysosomal acid lipases)
		abH14.2 (Gastric lipases)
GX	abH15 - Burkholderia lipases	abH15.1 (Staphylococcus aureus lipase like)
		abH15.2 (Burkholderia cepacia lipase like)
		abH15.3 (Saccharomyces cerevisiae lipase 2 like)
GX	abH16 - Streptomyces lipases	abH16.1 (Streptomyces lipases)
GX	abH17 - Chloroflexus aurantiacus lipase like	abH17.1 (Chloroflexus aurantiacus lipase like)

GX	abH18 - Bacillus lipases	abH18.1 (Bacillus lipases)
GX	abH19 - Thioesterases	abH19.1 (Palmitoyl-protein thioesterase 1 like)
		abH19.2 (Palmitoyl-protein thioesterase 2 like)
GX	abH20 - Lipoprotein lipases	abH20.1 (Hepatic lipases)
		abH20.2 (Lipoprotein lipases)
		abH20.3 (Pancreatic lipases)
GX	abH21 - Bacterial esterases	abH21.1 (Chlorobium esterases)
		abH21.2 (Rickettsia conorii esterase like)
		abH21.3 (Pesudomonas esterases)
		abH21.4 (Carboxylesterases)
		abH21.5 (Nostoc sp. esterase like)
GX	abH22 - Lysophospholipase	abH22.1 (Carboxylesterases)
		abH22.2 (Microbulbifer degradans esterase like)
		abH22.3 (Lysophospholipase)
		abH22.4 (Arabidopsis thaliana lysophosolipases)
		abH22.5 (Arabidopsis thaliana proteins)
		abH22.6 Trypanosoma brucei lysophospholipase like)
		abH22.7 (Homo sapiens lysophospholipase like)
		abH22.8 (Schizosaccharomyces pombe lysophospholipase like)
		abH22.9 (Saccharomyces cerevisiae proteins)
GX	abH23 - Filamentous fungi lipases	abH23.1 (Rhizomucor mihei lipase like)
		abH23.2 (Saccharomyces lipase like)
GX	abH24 - Pseudomonas lipases	abH24.1 (Pseudomonas lipases)
GX	abH25 - Moraxella lipase 1 like	abH25.1 (Moraxella lipase 1 like)
GX	abH26 - Deacetylases	abH26.1 (Deacetylases)
Y	abH27 - Dipeptidyl peptidase IV like	abH27.1 (Dipeptidyl peptidase IV like)
Y	abH28 - Prolyl endopeptidases	abH28.1 (Propyl endopeptidases)
Y	abH29 - Dipeptidyl-peptidases	abH29.1 (Dipeptidyl-peptidases)
Y	abH30 - Cocaine esterases	abH30.1 (Cocaine esterases)
GX	abH31 - Dienlactone Hydrolases	abH31.1 (Xanthomonas dienelactone hydrolases)
		abH31.2 (Carboxymethylenebutenolidases)
		abH31.3 (Caulobacter dienelactone hydrolases)
		abH31.4 (Nostoc dienelactone hydrolases)
		abH31.5 (Salmonella carboxymethylenebutenolidase)
		abH31.6 (Agrobacterium carboxymethylenebutenolidase)
		abH31.7 (Rhodococcus dienelactone hydrolases)
GX	abH32 - Xylanase esterases	abH32.1 (Xylanase Z esterase domain)
		abH32.2 (Xylanase Y esterase domain)
GX	abH33 - Antigen 85	abH33.1 (Antigen 85-C)
		abH33.2 (Antigen 85-A)
		abH33.3 (Antigen 85-B)
GX	abH34 - Lysosomal protective protein like	abH34.1 (Lysosomal protective protein like)
		abH34.2 (Serine carboxypeptidase II like)
GX	abH35 - Acyl-transferases	abH35.1 (Acyl-transferases)
GX	abH36 - Cutinases	abH36.1 (Colletotrichum cutinases)
		abH36.2 (Botryotinia cutinases)
		abH36.3 (Fusarium cutinases)
		abH36.4 (Mycobacterium cutinases)
GX	abH37 - Candida antarctica lipase like	abH37.1 (Candida antarctica lipase B like)

3.5.1 Konservierung innerhalb der α/β-Hydrolasen

Die Familie der α/β-Hydrolasen stellt eine der größten Proteinfamilien dar, wobei deren Mitglieder keine signifikante Sequenzähnlichkeit aufweisen müssen. Dennoch besitzen alle Vertreter dieser Familie ein gemeinsames Faltungsmotiv. Zur Bestimmung der für dieses Faltungsmotiv relevanten Aminosäuren wurde die Konservierung der Aminosäuren innerhalb acht repräsentativer Superfamilien untersucht (abH1, abH4, abH9, abH14, abH15, abH22, abH27 und abH33). Für die Auswahl der Familien war entscheidend, dass (1) eine Proteinstruktur innerhalb der Familie bekannt war und (2) mindestens 20 Sequenzen mit einer Sequenzidentität unter 90% vorlagen, da andernfalls der Informationsgehalt für die Konservierungsanalyse zu gering gewesen wäre.

Die mittlere Sequenzidentität und Sequenzähnlichkeit dieser Superfamilien ist in Abbildung 18 dargestellt. Es zeigte sich, dass diese zwischen den Superfamilien deutlich variieren kann. So reichte die Sequenzidentität von 22,9% für Superfamilie abH4 bis 41,1% für Superfamilie abH33.

Für die Berechnung der Konservierung wurde die *sum of pairs* Methode (SP) verwendet, die die physikalisch-chemischen Eigenschaften der Aminosäuren einbezieht. Die Abbildung der Konservierung auf Proteinstrukturen half bei der Visualisierung der Konservierung im dreidimensionalen Raum und unterstützte die Vorhersage der für die Faltung wichtigen Bereiche. Für die untersuchten Superfamilien abH1, abH4, abH9, abH14, abH15, abH22, abH27 und abH33 ergaben sich folgende mittlere Konservierungen C_m: 0,7, 0,7, 0,7, 1,5, 1,2, 1,7, 1,7 bzw. 2,6. Jeweils eine Superfamilie der drei verschiedenen Klassen GX (Superfamilie abH15), GGGX (Superfamilie abH1) und Y (Superfamilie abH27) werden hier genauer beschrieben.

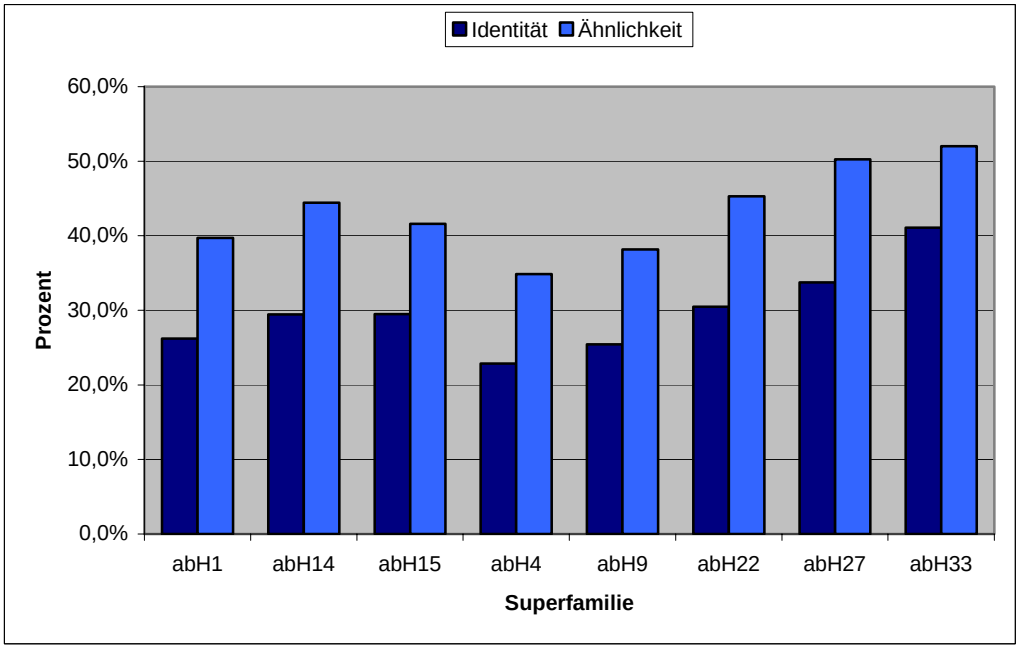


Abbildung 18: Darstellung der mittleren Sequenzidentität (dunkelblau) und Sequenzähnlichkeit (hellblau) für die acht ausgewählten Superfamilien, die in die Konservierungsanalyse einbezogen wurden.

In Abbildung 19 ist die Konservierung C pro Aminosäure der jeweiligen Superfamilie als Funktion der Proteinsequenz für die Referenzstrukturen 4LIP (abH15), 1AKN (abH1) und 1J2E (abH27) dargestellt.

Für die GX Superfamilie abH15 ist die Konservierung in Abbildung 19a dargestellt. Die Hydrolasen dieser Superfamilie zeichnen sich im Vergleich zum kanonischen α/β-Hydrolase Fold durch das Fehlen der β-Strands β1 und β2 aus. Es zeigte sich, dass Aminosäuren mit hoher Konservierung C überwiegend in Abschnitten lagen, die eine α- oder β-Sekundärstruktur ausbildeten. So waren die β-Strands β3, β4, β5 und β7 sowie die α-Helices αA, αB, αC und a9 im Vergleich zur mittleren Konservierung signifikant konserviert. Das Profil der Konservierung C zeigte, das diese innerhalb dieser Sekundärstrukturelemente oszilliert. Der N-terminale Bereich bis zur α-Helix αC innerhalb der Superfamilie war am höchsten konserviert. Dieser Bereich umfasste sowohl das *oxyanion hole* als auch das GXSXG Motiv in dem das katalytisch aktive Serin lokalisiert war. Der Bereich zwischen der α-Helix αC und dem katalytisch aktiven Aspartat wies keine spezifische Konservierung auf mit Ausnahme des β-Strands β7 sowie den Aminosäuren H114 und G116, die in einem Abschnitt ohne definierte Sekundärstruktur lokalisiert sind. Der C-terminale Bereich wies erneut eine erhöhte Konservierung um das katalytisch aktive Aspartat und Histidin auf, sowie die α-Helix αF, die den kanonischen α/β-Hydrolase Fold abschließt.

Die Konservierung C der Superfamilie abH15 wurde auf die Referenzstruktur 4LIP abgebildet (Abbildung 20). Dies zeigte, dass die Konservierung im Kern des Proteins deutlich erhöht war, während die peripheren Bereiche wie das Lid sowie die dem Lösungsmittel zugänglichen Aminosäuren keine Konservierung aufwiesen. Die Abbildung verdeutlich auch die oszillierenden Konservierungsprofile innerhalb der α-helikalen Sekundärstrukturelemente. Während die äußeren lösungsmittelzugänglichen Positionen nicht konserviert waren, wiesen die in den Kern des Proteins gerichteten Aminosäuren eine signifikante Konservierung auf.

In Abbildung 19b ist für die GGGX Superfamilie abH1 die Konservierung als Funktion der Aminosäurepositionen in der Referenzstruktur 1AKN dargestellt. Die Hydrolasen dieser Superfamilie besitzen ein zentrales β-Faltblatt, das im Vergleich zum kanonischen α/β-Hydrolase Fold sowohl N- als auch C-terminal erweitertet ist. Das Profil der Konservierung C innerhalb der Superfamilie abH1 erwies sich als ähnlich zu dem der Superfamilie abH15. Die Aminosäuren mit hoher Konservierung lagen überwiegend in Abschnitten die eine α- oder β-Sekundärstruktur ausbildeten. So waren die β-Strands β-2, β-1, β1, β2, β3, β4, β5, β6, β7 und β8 sowie die α-Helices αB, αC, $\alpha D'_2$, αE, $\alpha F'_1$, $\alpha F'_2$ und αF im Vergleich zur mittleren Konservierung signifikant konserviert. Auch hier zeigte sich für diese Sekundärstrukturelemente ein oszillierendes Profil der Konservierung C. Der N-terminale Bereich war ebenfalls deutlich höher konserviert und erstreckte sich sogar bis β-Strand β6. Der Bereich zwischen β6 und dem C-Terminus wies eine Konservierung C deutlich unterhalb der mittleren Konservierung auf, mit Ausnahme für die Bereiche um das katalytische Aspartat und Histidin, die Sekundärstrukturelemente β7, β8, $\alpha D'_2$, αE, $\alpha F'_1$, $\alpha F'_2$ und αF sowie die Positionen P300 und Y367, die in einem Bereich ohne definierte Sekundärstruktur lokalisiert sind. Die Konservierung der Superfamilie abH1 wurde auf die Referenzstruktur 1AKN abgebildet (Abbildung 21). Auch für diese Superfamilie, deren Mitglieder im Schnitt 200 Aminosäuren länger sind als die der Superfamilie abH15, zeigte sich, dass der Kern des Proteins deutlich höher konserviert war, als die lösungsmittelzugänglichen Aminosäuren. Die Insertierungen, die zu der Verlängerung dieser Proteine führen, sind überwiegend peripher lokalisiert und wiesen ebenfalls keine signifikante Konservierung auf.

In Abbildung 19c ist die Konservierung C für die Y Klasse Superfamilie abH27 als Funktion der Aminosäureposition der Referenzstruktur 1J2E dargestellt. Diese Superfamilie wies eine relativ geringe mittlere Sequenzlänge auf, wobei der kanonische α/β-Hydrolase Fold jedoch vollständig erhalten war. Auch für diese Superfamilie wurde beobachtet, dass der Grossteil der signifikant konservierten Aminosäuren in Abschnitten mit oszillierendem

Konservierungsprofil lagen, die eine α- oder β-Sekundärstruktur ausbildeten. So waren die β-Strands β3, β4, β5, β6, β7, und β8 sowie die α-Helices αB, αC, $\alpha D'_{1,}$ αE und αF im Vergleich zur mittleren Konservierung signifikant konserviert. Neben diesen Bereichen wurden deutlich mehr Aminosäuren mit hoher Konservierung beobachtet, die in Abschnitten ohne definierter Sekundärstruktur lokalisiert waren (P531, P550, G582, G587, P655, W659). Im Unterschied zu den Superfamilien abH1 und abH15 zeigte die Superfamilie abH27 auch keinen größeren Bereich mit geringer Konservierung C auf. Der in den zuvor untersuchten Superfamilien beschriebene Abschnitt zwischen αC und αE mit geringer Konservierung war in dieser Superfamilie um etwa 100 Aminosäuren kürzer und somit auf die Sekundärstrukturelemente des kanonischen α/β-Hydrolase Folds beschränkt. Die Konservierung der Superfamilie abH27 wurde auf die Referenzstruktur 1J2E abgebildet (Abbildung 22). Für diese Superfamilie zeigte sich, dass auch hier überwiegend die für das Lösungsmittel nicht zugänglichen Aminosäuren im Kern des Proteins hoch konserviert waren.

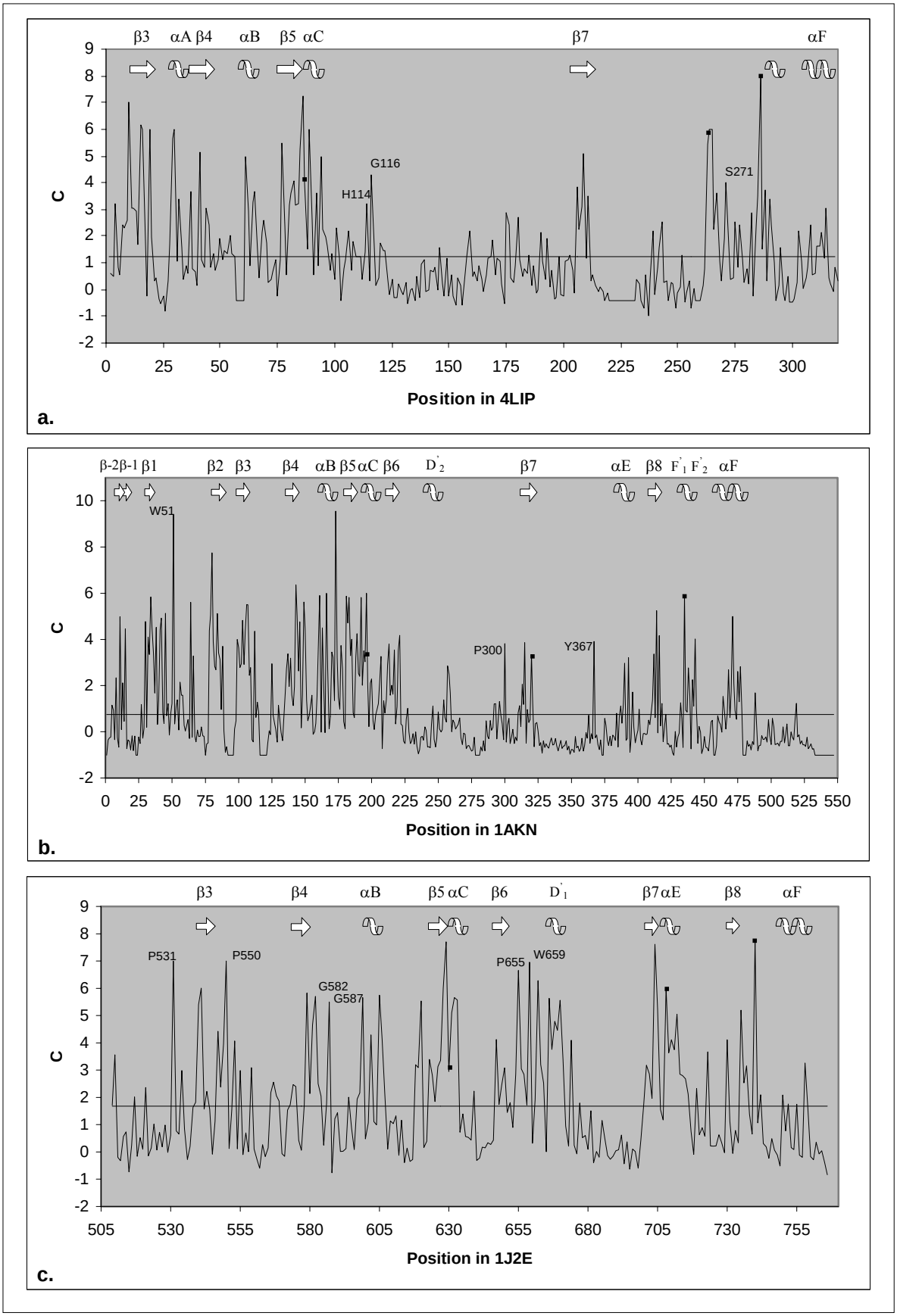


Abbildung 19: Auftragung der Konservierung C innerhalb einer Superfamilie gegen die Positionen einer Referenzstruktur: a. Superfamilie abH15, b. Superfamilie abH1, c. Superfamilie abH27. α-Helices und β-Strands sind durch entsprechende Piktogramme, die Aminosäuren der katalytischen Triade durch Punkte markiert. Aminosäuren mit hoher Konservierung C außerhalb von Sekundärstrukturbereichen sind beziffert.

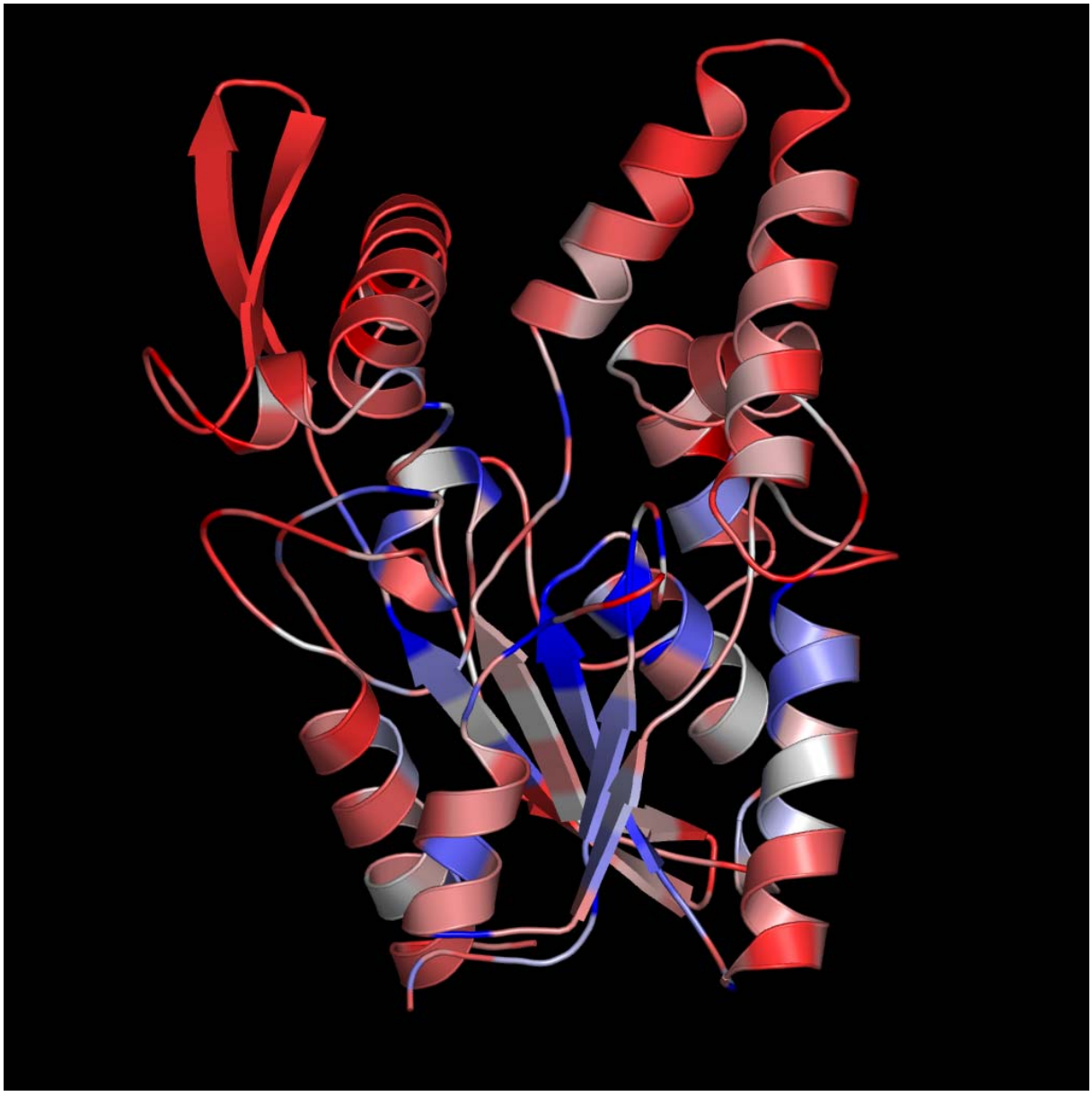


Abbildung 20: Abbildung der Konservierung C der GX-Klasse Superfamilie abH15 auf die Referenzstruktur 4LIP der Lipase aus *Burkholderia cepacia*. Das Spektrum der Konservierung C verläuft von rot (schwach konserviert) bis blau (stark konserviert).

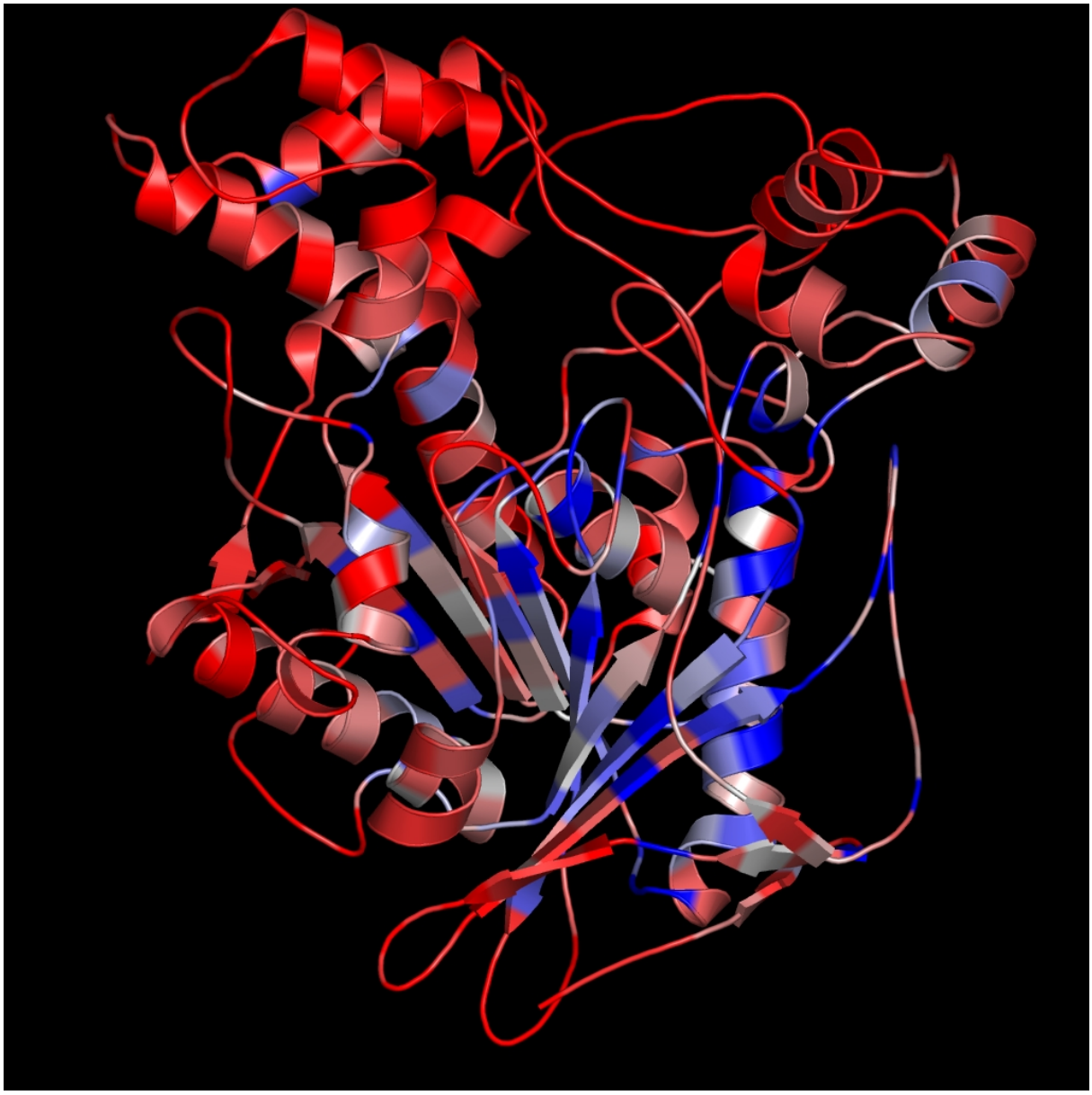


Abbildung 21: Abbildung der Konservierung C der GGGX-Klasse Superfamilie abH1 auf die Referenzstruktur 1AKN der Bile-salt aktivierten Lipase aus *Bos taurus*. Das Spektrum der Konservierung C verläuft von rot (schwach konserviert) bis blau (stark konserviert).

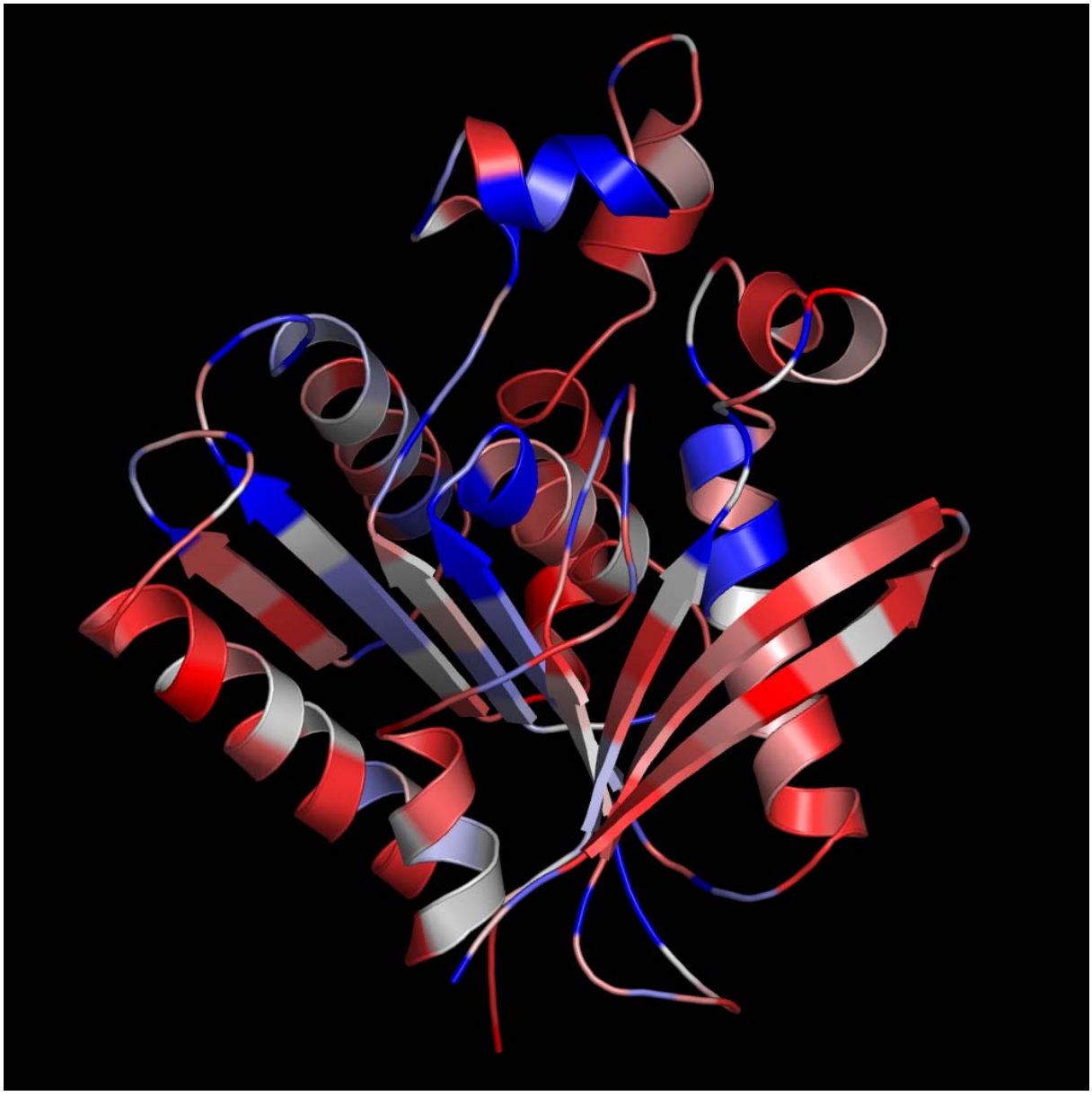


Abbildung 22: Abbildung der Konservierung C der Y-Klasse Superfamilie abH27 auf die Referenzstruktur 1J2E der Dipeptidyl Peptidase IV aus *Homo sapiens*. Das Spektrum der Konservierung C verläuft von rot (schwach konserviert) bis blau (stark konserviert).

3.5.2 Vergleich der repräsentativen Strukturen

Da die Überlagerung der in der LED abgelegten Strukturen auf den Vergleich der Bindungstasche und nicht eine globale Überlagerung des gesamten Folds optimiert war, musste für die Analyse des Folding Cores die Überlagerung angepasst werden. Die 8 ausgewählten Strukturen 1AKN (abH1), 1FJ2 (abH22), 1K8Q (abH14), 1LZL (abH4), 1QO7 (abH9), 4LIP (abH15), 1F0N (abH33) und 1J2E (abH27) ließen sich mit einem maximalen RMSD von 2,97 Å für die 29 zur Überlagerung gewählten Atome zwischen den PDB Einträgen 4LIP und 1QO7 überlagern (Tabelle 3). Der minimale RMSD ergab sich mit 0,56 Å für die Überlagerung der PDB Einträge 1J2EA und 1FJ2. Aus dem für diese Überlagerung erstellten und manuell überarbeiteten Strukturalignment (Abbildung 23) ergaben sich so insgesamt 106 Positionen die strukturell erhalten waren. Hiervon waren 75% in den β-Strands β3 bis β8 und den α-Helices αB, αC, αE und αF lokalisiert. Am strukturell unähnlichsten waren die Regionen zwischen β4 und αB sowie β6 und β7, die durch Insertionen unterschiedlichster Länge innerhalb der einzelnen Strukturen nicht vergleichbar waren. Auch die einzelnen Sekundärstrukturelemente des kanonischen α/β-Hydrolase Folds wiesen bereits deutliche Unterschiede auf. So variierte z.B. die Länge des β-Strands β5 nach der in PDBsum[99] vorgegebenen Annotationen von 5 Aminosäuren in 1QO7A bis 12 Aminosäuren in 1J2EA bzw. reichte die Länge der α-Helix αE von 4 Aminosäuren in 4LIPD bis 17 Aminosäuren in 1FJ2. Die größte Variabilität der Sekundärstrukturelemente wies jedoch die Region um die α-Helix αD auf. Diese ist weder in der Lage noch in der Länge erhalten.

Tabelle 3: RMSD Werte für Überlagerung der 8 Referenzstrukturen auf 1AKN

	1AKN	1FJ2	1K8QB	1LZL	1QO7A	4LIPD	1F0NA	1J2EA
1AKN	0,0	1,43	1,47	1,56	2,14	1,99	1,11	1,33
1FJ2		0,0	1,11	0,82	2,52	2,27	1,27	0,56
1K8QB			0,0	1,50	2,63	2,47	1,00	0,94
1LZL				0,0	2,57	2,20	1,40	0,94
1QO7A					0,0	2,97	2,51	2,62
4LIPD						0,0	2,25	2,26
1F0NA							0,0	1,00
1J2EA								0,0


β0 β1
1AKN 1 AKLGSVYTE GGFVEGVNKK LSLFGDSIDI FKGIPFAAAP KALEKPERHP
1FJ2
1K8QB 1 AFGKLHP TNPEVTMNIS QMITYWGYPA EEYEVVT--- ----------
1LZL 2 TTFPTLDPEL AAALTMLPKV DFADLPNARA TYDALIGAML ADLSFDGVSL RE-LSAPGLD G---------
1QO7A 3 KAFAKFP
4LIPD
1F0NA 2 SRPGLPV EYLQVPSPSM ----------
1J2EA 509 MPS KKLDFIILN- ----------
1AKN 50 GWQGTLKAKS FKKRCLQATL TQDS------ ---------- ---------- ---------- ----------
1FJ2
1K8QB ---------- ---------- ---------- ---------- ---------- ---------- ----------
1LZL 61 ---------- ---------- ---------- ---------- ---------- ---------- ----------
1QO7A 84 SSASISPNPF TVSIPDEQLD DLKTLVRLSK IAPPTYESLQ ADGRFGITSE WLTTMREKWL SEFDWRPFEA
4LIPD
1F0NA ---------- ---------- ---------- ---------- ---------- ---------- ----------
1J2EA ---------- ---------- ---------- ---------- ---------- ---------- ----------
β2 β3
1AKN 74 ---------T YGNEDCLYL- NIWVPQGRK- EVSHDLPVMI WIY-GGAFLM GASQGANFLS NYLYD----G
1FJ2 1 MSTPLP- AIVPAAR--- ---KATAAVI FLH-GLGD-- ---------- ------T GHGWAEAF--
1K8QB 35 ---------- -EDGYILGI- DRIPYGRKNS ENIGRRPVAF LQH-GLLASA TNWI------ --SNLPNNSL
1LZL 62 ---------- ---DPEVKI- RFVTPDNT-- --AGPVPVLL WIH-GGGFAI GTAES----- ---SDP--FC
1QO7A 80 RLNSFPQFTT EIEGLTIHFA ALFSE----- --REDAVPIA LLH-GWPGSF VEFYPIL--- ------QLFR
4LIPD 2 D NYAATRYPII LVH-GL---T GTD-KYAGVL EY---WYGIQ
1F0NA 19 ---------- ---GRDIKV- QFQSGG---- ---NNSPAVY LLD-GL---R AQDD------ YNGWDINTPA
1J2EA 521 ---------- --ETKFWYQ- MILPPHFD-- -KSKKYPLLL DVYAGPCSQK ---------- ADTVFRLNWA
αA β4
1AKN 128 EEI-ATRGN- VIVVTFNYRV G--------- -------PLG FL-------- ------STGD SNLPGNY---
1FJ2 37 ---AGIRSSH IKYICPHAPV RPVTLNMNVA MP-------- --------SW FDIIGLSPDS QEDESGIKQ-
1K8QB 84 AFILAD-AG- YDVWLGNSRG NTWARRNLYY SPDSVEFWAF SF-------- ---------- -----DE---
1LZL 103 VEV-ARELG- FAVANVEYRL A--------- -------PET T--------- ---------- ----FPG---
1QO7A 133 EEYTPETLP- FHLVVPSLPG YTFSS----- ---------- ---------- ---------- GPPLDKDFGL
4LIPD 35 EDL-QQR-G- ATVYVANLSG ---------- ---------- ---------- ---------- FQSDDGPNG-
1F0NA 58 FE-WYYQS-G LSIVMPVGGQ ---------- ---------- ---------- ---------- ------SSFY
1J2EA 565 TYLAST--EN IIVASFDGRG SG-------- -YQGDKIMHA INRRLGT--- ---------- ----------
αB β5 * αC
1AKN 163 ---------- ---GLWD-QH MAIAWVK-RN ---IEAFGGD PD-NITLFGE SAGGASVSLQ TLSPYN----
1FJ2 87 ---------- ---AAEN-IK ALIDQEV-KN GI-------P SN-RIILGGF SQGGALSLYT ALT--T----
1K8QB 127 ---------- ---MAKYDLP ATIDFIL-KK ---TGQD--- ---KLHYVGH SQGTTIGFIA FSTNPK----
1LZL 129 ---------- ---PVND-CY AALLYIH-AH AEELG---ID PS-RIAVGGQ SAGGGLAAGT VLKARDE---
1QO7A 167 MD-------- ---NARV-VD QLMKDLG--- ---------F GS-GYIIQGG DIGSFVGRLL GVGF------
4LIPD 62 ---------- ---RGEQ-LL AYVKTVLAA- --------TG AT-KVNLVGH SQGGLTSRYV --AAVA----
1F0NA 80 SDWYSPACGK AGCQTYK-WE TFLTSELPQW LSAN--RAVK PT-GSAAIGL SMAGSSAMIL AAY-------
1J2EA 601 ---------- --FEVED-Q- IEAARQFSKM G-------FV DNKRIAIWGW SYGGYVTSMV LGSG------
β6
1AKN 210 -KGLIKRAIS QSGVGLCPWA IQQDPLFWAK RIAEKVGCPV DDTSKMAGCL KITDP----- --------RA
1FJ2 128 -QQKLAGVTA LSCWLPLRAS FP-------- ---------- ---------- ---------- ----------
1K8QB 169 LAKRIKTFYA LAP--VATVK YTETLINKLM LVPSFLFKLI FGNKIFYPHH FFDQFLATEV CSRETVDLLC
1LZL 177 GVVPVAFQFL EIPE-LDDRL ETVSMTNF-- ---------- ---------- ---------- ----------
1QO7A 206 --DACKAVHL NLCAMRAPPE GPSIESLSAA EKEGIARMEK FMTDGLAYAM EHSTRPSTIG HVLSS-----
4LIPD 101 -PDLVASVTT IGTP-HRGSE ---------- ---------- ---------- ---------- ----------
1F0NA 139 HPQQFIYAGS LSALLDPSQG MGPSLIGLAM GDAGGY---- ---------- ---------- ----------
1J2EA 644 -SGVFKCGIA VAP--VSRWE ---------- ---------- ---------- ---------- ----------
1AKN 266 L---------T LAYKLPLGST EYPKLH---- ---------- ---------- ---------- ----YL---S
1FJ2 ---------- ---------- ---------- ---------- ---------- ---------- ----------
1K8QB 237 SNALFIICG- ---------- ---------- ---------- ---------- ---------- ----------
1LZL 204 ---------- ---------- ---------- ---------- ---------- ---VDTPLWH RPNAILS-WK
1QO7A ---------- ---------- ---------- ---------- ---------- ---------- ----------
4LIPD 119 ---------- ---------- FADFVQGVLA YDPTGLSSTV IAAFVNVFGI LTSSSNNTNQ DALAAL---K
1F0NA ---------- ---------- ---------- ---------- ---------- ---------- ----------
1J2EA 661 ---------- ---------- ---YYD---- ---------- ---------- ---------- ----SVYTER
1AKN 287 FVPVIDGDF- ---------- ---------- ---------- ---------- ---------- ----------
1FJ2 ---------- ---------- ---------- ---------- ---------- ---------- ----------
1K8QB 246 ---------- ------FDTM NLNMSRLDVY LSHNPAGTSV QNVLHWSQAV KSGKFQAFDW GSPVQNMMHY
1LZL 220 YYLGESYSGP EDPDVS---- ---------- ---------- ---------- ---------- ----------
1QO7A 269 ---------- ---------- -------SPI ALLAWIGEKY LQWVDKPLPS ETILEMVSLY WLTESFPRAI
4LIPD 166 TLTTAQA--- ---------- ---------- ---------- ---------- ---------- ----------
1F0NA 175 ---------- ---------- ---------- ---------- ---------- ---------- -KAADMWGPS
1J2EA 670 YMGLPTPEDN LDHY------ ---------- ---------- ---------- ---------- ----------

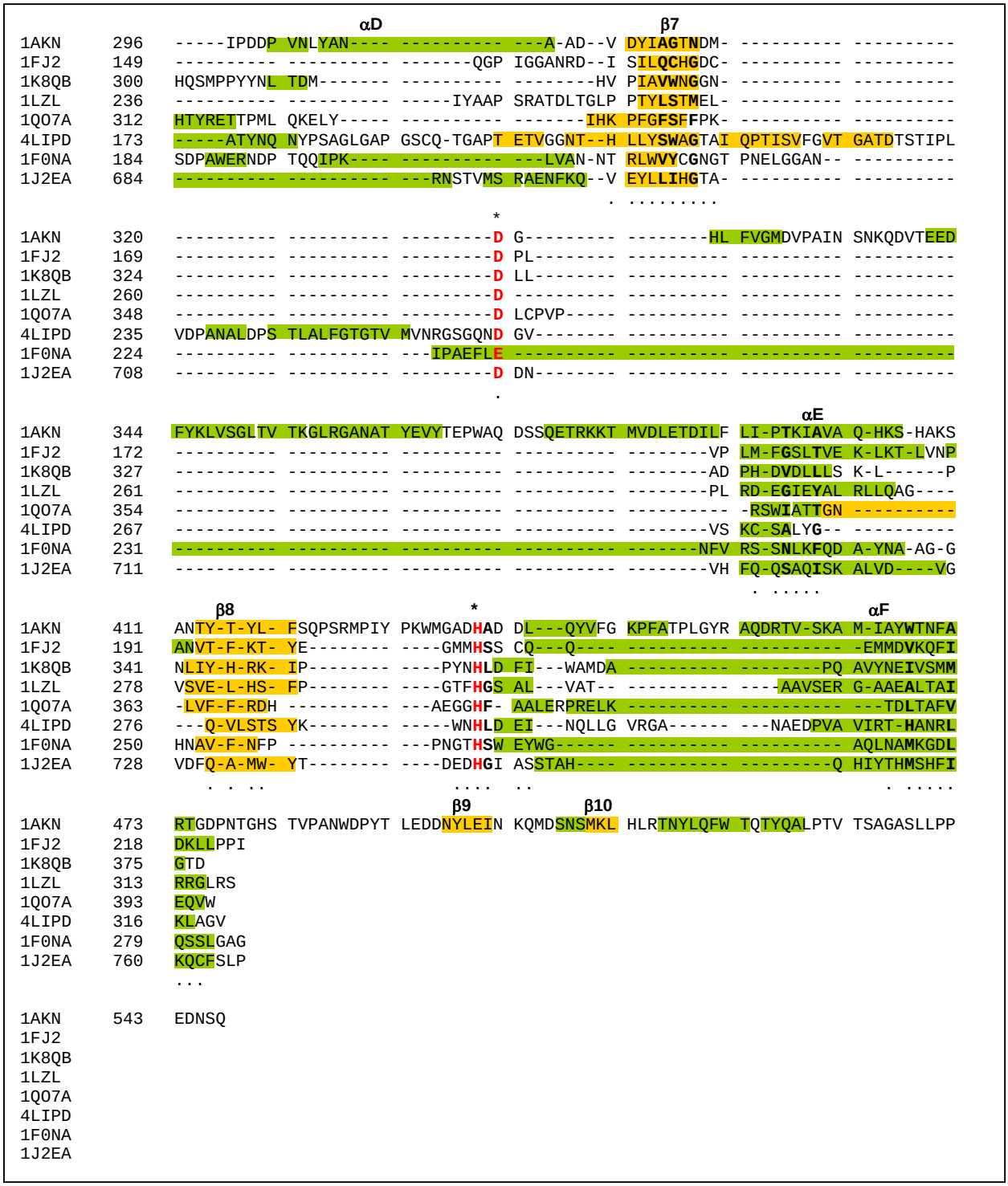


Abbildung 23: Strukturalignment der acht Referenzstrukturen 1AKN (abH1), 1FJ2 (abH22), 1K8Q (abH14), 1LZL (abH4), 1QO7 (abH9), 4LIP (abH15), 1F0N (abH33) und 1J2E (abH27). α-Helices sind orange markiert, β-Strands grün. Die Sekundärstrukturelemente des α/β-Hydrolase Folds sind beschriftet. Alle im Strukturalignment übereinstimmenden Positionen sind mit einem Punkt markiert. Streng konservierte Positionen (SKPs) sind fett dargestellt. Die Aminosäuren der katalytischen Triade sind in rot dargestellt und mit einem Stern versehen.

3.5.3 Zwingend konservierte Positionen (ZKPs)

Insgesamt wurden 13 ZKPs identifiziert (Tabelle 4), deren Konservierung C in Abbildung 24 dargestellt ist. Neben den Aminosäuren der katalytischen Triade zeigten auch weitere Positionen diese hohe Konservierung. Die das katalytische Nucleophil N (ZKP_8) umgebenden Positionen N_{-2}, N_{-1} und N_{+1} sind Teil des Serinhydrolasemotivs GX_1SX_2G, das den *nucleophilic elbow* ausbildet. Jedoch wurde nicht die Position N_{+2} als ZKP identifiziert sondern der C-terminale Nachbar N_{+3}. N_{+2} stellt keine ZKP dar, da für die Superfamilie abH33 an dieser Position neben dem eigentlich erwarteten Glycin auch Alanin und Serin zu beobachten waren und so die Konservierung C unter den Durchschnittswert fiel. Für die variablen Positionen X_1 (ZKP_7) und X_2 (ZKP_9) des Serinhydrolasemotivs zeigte sich, dass diese zwar innerhalb der Superfamilien hoch konserviert waren, die Eigenschaften zwischen den Superfamilien jedoch stark variierten, wobei Position X_2 als tendenziell hydrophob eingestuft werden konnte. Des Weiteren wurden die Aminosäuren des HG Motivs N-terminal zur *oxyanion hole* formenden Aminosäure als ZKP_1 und ZKP_2 bestimmt. Neben diesen Positionen wurde noch 4 weitere Positionen als ZKPs identifiziert. ZKP_3 und ZKP_5 sind beides Positionen im hydrophoben Kern des Proteins und nicht zugänglich für das Lösungsmittel. ZKP_3 ist im β-Strand β4 lokalisiert, ZKP_5 ist Teil der α-Helix αB. ZKP_{11} ist 3 Positionen N-terminal zur katalytisch aktiven Säure am Ende des β-Strands β7 lokalisiert. Das hoch konservierte Glycin an dieser Position ist dadurch zu erklären, dass β7 und die katalytisch aktive Säure durch einen scharfen Knick verbunden sind, dessen Ausbildung durch das Glycin erleichtert wird. Die bisher beschriebenen ZKPs sind aufgrund der Funktion der Hydrolasen, der gesonderten Stellung des Glycins in engen Turns oder der Ausbildung des für Proteine typischen hydrophoben Kerns zu erwarten gewesen. Eine Ausnahme stellte ZKP_4 dar. Lokalisiert in einem Loop C-terminal zum β-Strand β4 war in 4 von 8 Fällen (abH1, abH14, abH4, abH27) an dieser Position ein Arginin hoch konserviert (99%, 100%, 80%, 98%). Für die Superfamilie abH22 war ein hochkonserviertes Prolin zu finden (88%). Superfamilie abH9 stellte mit vergleichbaren Anteilen für Arginin (45%) und Prolin (52%) einen Übergang zwischen diesen beiden Zuständen dar. In den beiden verbleibenden Superfamilien abH33 und abH15 fanden sich an dieser Position ein hoch konserviertes Glycin (90%) bzw. ein relativ hoch konserviertes Serin (60%). Die Untersuchung der repräsentativen Strukturen mit konserviertem Arginin ergab, dass dieses immer mit einem Aspartat wechselwirkte (1AKN: D166, 1K8Q: D130, 1LZL: D132, 1J2E: D130). Der Abstand der Schwerpunkte zwischen den geladenen Gruppen des Arginins und Aspartats lag hierbei

zwischen 3,8 Å und 4,3 Å und somit in dem für die Ausbildung einer Salzbrücke notwendigen Bereich. Die Bestimmung der Konservierung innerhalb der Superfamilien mit konservierten Arginin zeigte, dass auch das wechselwirkende Aspartat innerhalb der Superfamilien hoch konserviert war (Superfamilie abH1: 100%, abH14: 100%, abH4: 79%, abH27: 98%).

Tabelle 4: Liste der 13 zwingend konservierten Positionen (ZKPs) der acht untersuchten Superfamilien.

	abH15	abH9	abH14	abH33	abH22	abH1	abH4	abH27	Rolle
ZKP_1 (SKP_5)	H(90%)	H(84%)	H(96%)	D(85%)	H(100%)	H(62%)	H(95%)	Y(85%)	*Oxyanion Hole* Motiv
ZKP_2 (SKP_6)	G(100%)	G(84%)	G(96%)	G(100%)	G(100%)	G(96%)	G(95%)	G(85%)	*Oxyanion Hole* Motiv
ZKP_3 (SKP_{10})	V(87%)	V(49%)	V(96%)	V(75%)	F(33%), V(9%)	V(60%)	V(74%)	V(70%)	Hydrophober Kern
ZKP_4 (SKP_{16})	S(60%)	P(52%), R(45%)	R(100%)	G(90%)	P(88%)	R(99%)	R(80%)	R(98%)	Stabilisierung des β-Strands β4 und der α-Helix αB
ZKP_5 (SKP_{20})	L(97%)	I(28%)	L(79%)	W(100%)	V(52%)	Q(80%)	C(42%)	Q(93%)	Hydrophober Kern
ZKP_6 (SKP_{28})	G(93%)	G(90%)	G(90%)	G(100%)	G(100%)	G(98%)	G(100%)	G(100%)	Erstes Glycin des GXSXG Motivs
ZKP_7 (SKP_{29})	H(93%)	G(79%)	H(81%)	L(70%)	F(94%)	E(56%)	D(59%)	W(87%)	Erstes X des GXSXG Motivs
ZKP_8 (SKP_{30})	S(100%)	S(63%)	S(96%)	S(100%)	S(100%)	S(94%)	S(100%)	S(85%)	Katalytisch aktives Serin
ZKP_9 (SKP_{31})	Q(50%)	W(72%)	Q(88%)	M(100%)	Q(73%)	A(87%)	A(74%)	Y(85%)	Zweites X des GXSXG Motivs
ZKP_{10} (SKP_{33})	G(63%)	S(67%)	T(73%)	G(80%)	G(76%)	G(47%)	G(85%)	G(98%)	*Nucleophilic Elbow*
ZKP_{11} (SKP_{47})	G(80%)	G(63%)	G(62%)	G(95%)	G(94%)	N(45%)	A(57%)	G(93%)	Haarnadelschleife
ZKP_{12} (SKP_{48})	D(100%)	D(72%)	D(100%)	E(95%)	D(97%)	E(83%)	D(100%)	D(100%)	Katalytisch aktive Säure
ZKP_{13} (SKP_{51})	H(100%)	H(99%)	H(100%)	H(100%)	H(97%)	H(88%)	H(98%)	H(98%)	Katalytisch aktives Histidin

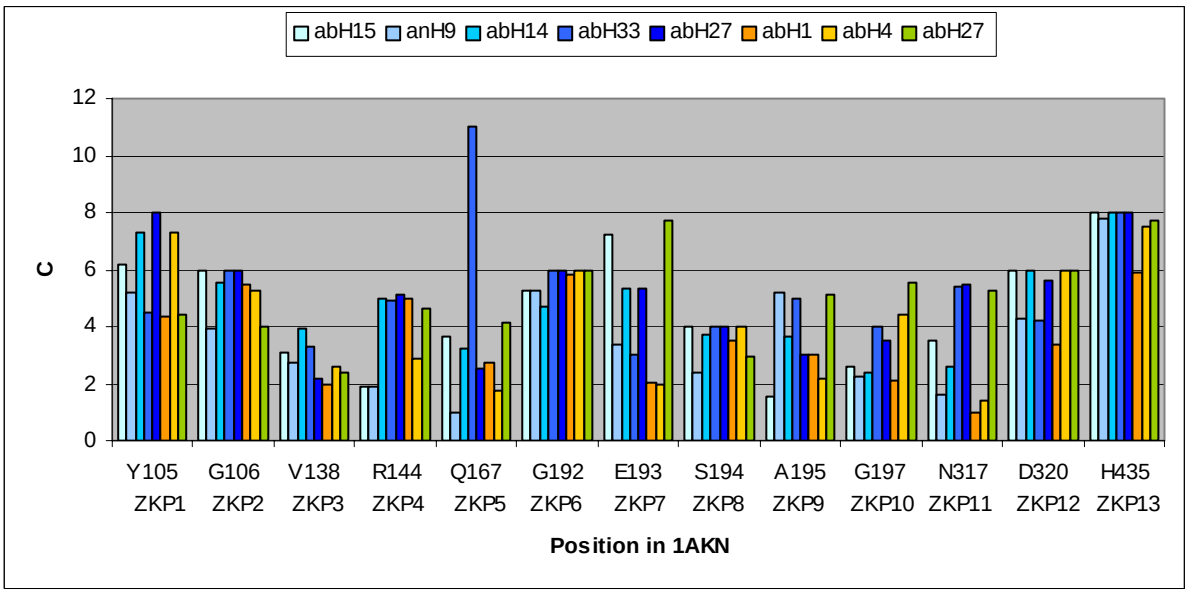


Abbildung 24: Konservierung C der zwingend konservierten Positionen für die acht untersuchten Superfamilien aufgetragen gegen die Aminosäurenummerierung der Referenzstruktur 1AKN der Superfamilie abH1. Superfamilien der GX-Klasse sind in blau dargestellt, Superfamilien der GGGX-Klasse in gelb, Superfamilie der Y-Klasse in grün.

3.5.4 Streng konservierte Positionen (SKPs)

Insgesamt wurden 54 der 106 alignten Positionen als SKPs mit einer Konservierung C größer Null identifiziert (Anhang SKPs). Die Konservierung C der SKPs ist in Abbildung 25 dargestellt. Die Verteilung der Aminosäurehäufigkeit innerhalb der streng konservierten Positionen wurde mit der Verteilung innerhalb der LED90 verglichen. Hierbei wurde jedoch die katalytischen Triade nicht in die Verteilung der streng konservierten Positionen mit einbezogen, da diese Aminosäuren durch die für die chemische Reaktion nötigen Eigenschaften bestimmt werden, und nicht aufgrund der Fähigkeit, die Proteinstruktur zu stabilisieren. Der Vergleich zeigte einen signifikante Unterschied. Wie in Abbildung 26 dargestellt wurde für Glycin, Histidin Isoleucin, Leucin, Methionin, Valin und Tryptophan festgestellt, dass diese innerhalb der streng konservierten Positionen deutlich überrepräsentiert waren im Vergleich zu ihren Hintergrundhäufigkeit in der LED90. Die hohe Häufigkeit des Glycins lässt sich darauf zurückführen, dass diese Aminosäuren in vielen Fällen für strukturell überlagerte Position unverändert vorlag. Diese Regionen sind überwiegend die Bereiche des *oxyanion hole* oder des *nucleophilic elbow*, die sterisch anspruchsvolle Aminosäuren ausschließen. Für Isoleucin, Leucin Methionin und Valin war zu beobachten, dass diese häufig gemeinsam eine streng konservierte Positionen dominieren und selten diese Positionen alleine besetzen. Trotz dessen, dass Prolin und Arginin innerhalb der streng konservierten Positionen deutlich unterrepräsentiert waren, fanden sich Positionen in einzelnen Superfamilien, wo diese vollständig an einer Position erhalten waren. Dies war im Falle von Arginin auf die Ausbildung von Salzbrücken zurückzuführen.

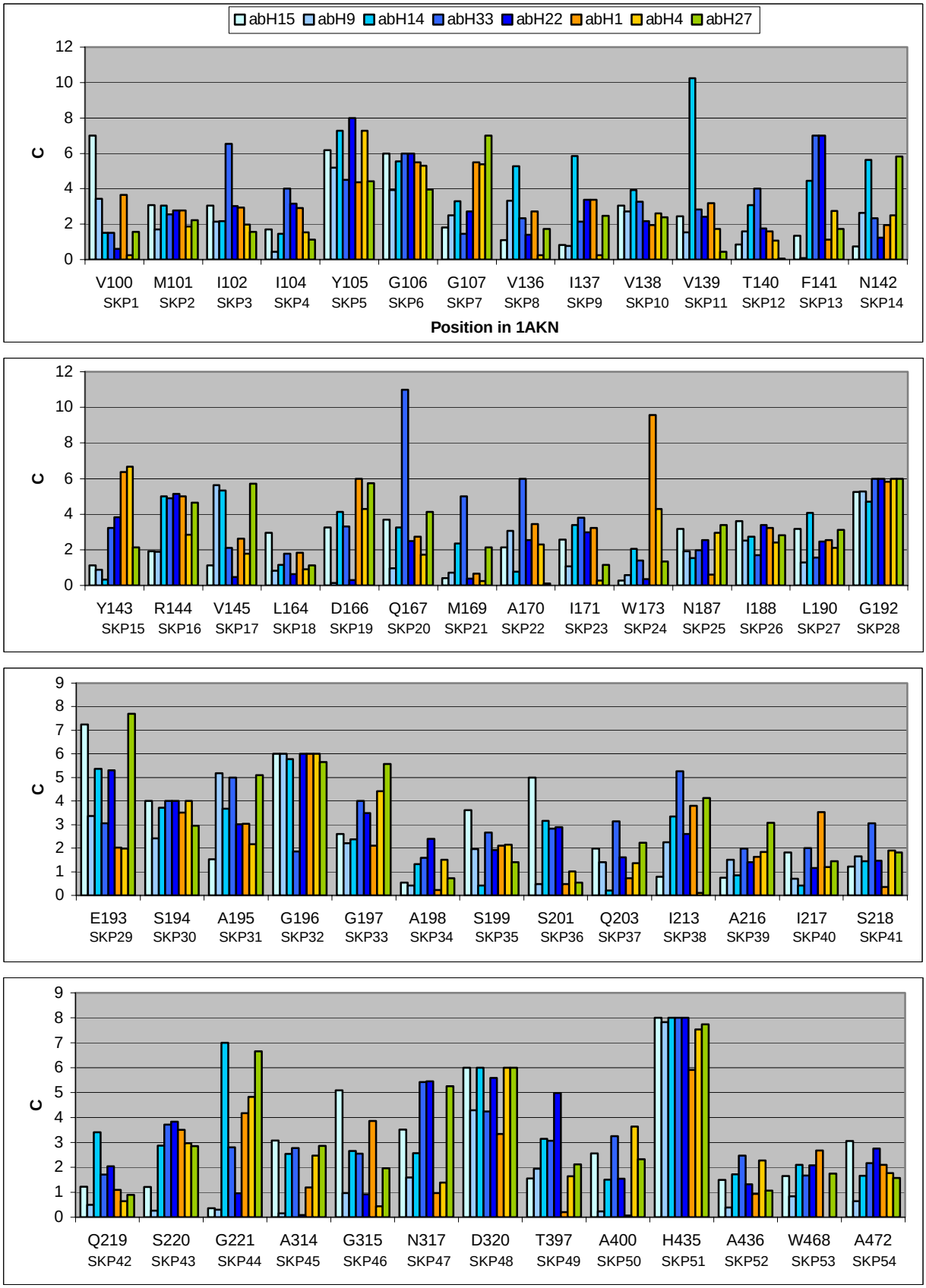


Abbildung 25: Konservierung C der streng konservierten Positionen (SKPs) der acht untersuchten Superfamilien aufgetragen gegen die Aminosäurenummerierung der Referenzstruktur 1AKN. Superfamilien der GX-Klasse sind in blau dargestellt, Superfamilien der GGGX-Klasse in gelb, Superfamilie der Y-Klasse in grün.

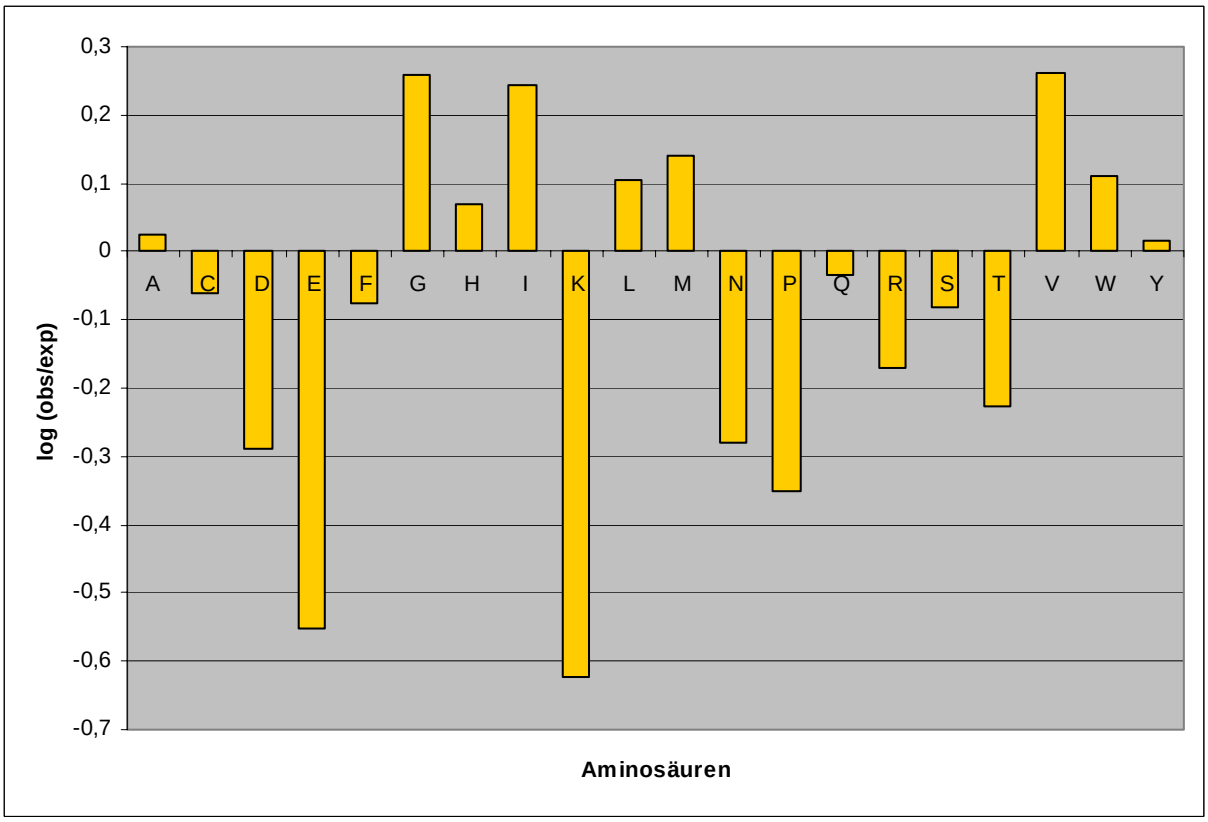


Abbildung 26: Verteilung der Aminosäuren in den streng konservierten Positionen. Logarithmische Auftragung des Verhältnis der Aminosäurehäufigkeit der streng konservierten Positionen (obs) ohne die Aminosäuren der katalytischen Triade zu der Aminosäurehäufigkeit innerhalb der LED90 (exp).

Für 9 SKPs waren in mindestens 50% der untersuchten Referenzstrukturen mehr als 5% der relativen Aminosäureoberfläche für das Lösungsmittel zugänglich (SKP_7, SKP_{16}, SKP_{19}, SKP_{21}, SKP_{24}, SKP_{25}, SKP_{30}, SKP_{51}, SKP_{52}). Der erste *oxyanion hole* Bildner (SKP_7), das katalytisch aktiven Nucleophil (SKP_{30}) und Histidin (SKP_{51}) sowie dessen C-terminaler Nachbar (SKP_{52}) sind funktionell bedingt für das Lösungsmittel zugänglich. Neben diesen Positionen waren 3 für das Lösungsmittel zugängliche SKPs (SKP_{19}, SKP_{21}, SKP_{24}) in der α-Helix αB und SKP_{25} im N-terminalen Bereich des β-Strands β5 lokalisiert. Die neunte Aminosäure war das als ZKP_4 identifizierte und eine Salzbrücke ausbildende Arginin (SKP_{16}). SKP_{19} stellte den Salzbrückenpartner für SKP_{16}. SKP_{21} war in 6 Familien zugänglich (abH15, abH14, abH33, abH9, abH22, abH1) und war entweder durch ein Alanin oder eine hydrophile Aminosäure besetzt mit Ausnahme der Familie abH27, in der diese Position streng hydrophob und nicht für das Lösungsmittel zugänglich war sowie der Familie abH4, die zwar ein Alanin bzw. Aspartat an dieser Position trug, jedoch nicht für das Lösungsmittel zugänglich war. SKP_{24} war ebenfalls in 6 Familien zugänglich (abH15, abH4, abH14, abH27,

abH33, abH22). In 3 Familien (abH4, abH14, abH1) war diese Position mit sterisch anspruchsvollen, hydrophoben Aminosäuren besetzt (W, F, Y) während in den restlichen Familien hydrophile Aminosäuren zu finden waren (T, S, K, R). SKP_{25} war mit Ausnahme von abH9 in allen Familien für das Lösungsmittel zugänglich. Für jede Familie zeigte sich auch eine deutliche Konservierung eines Arginins oder Lysins an dieser Position.

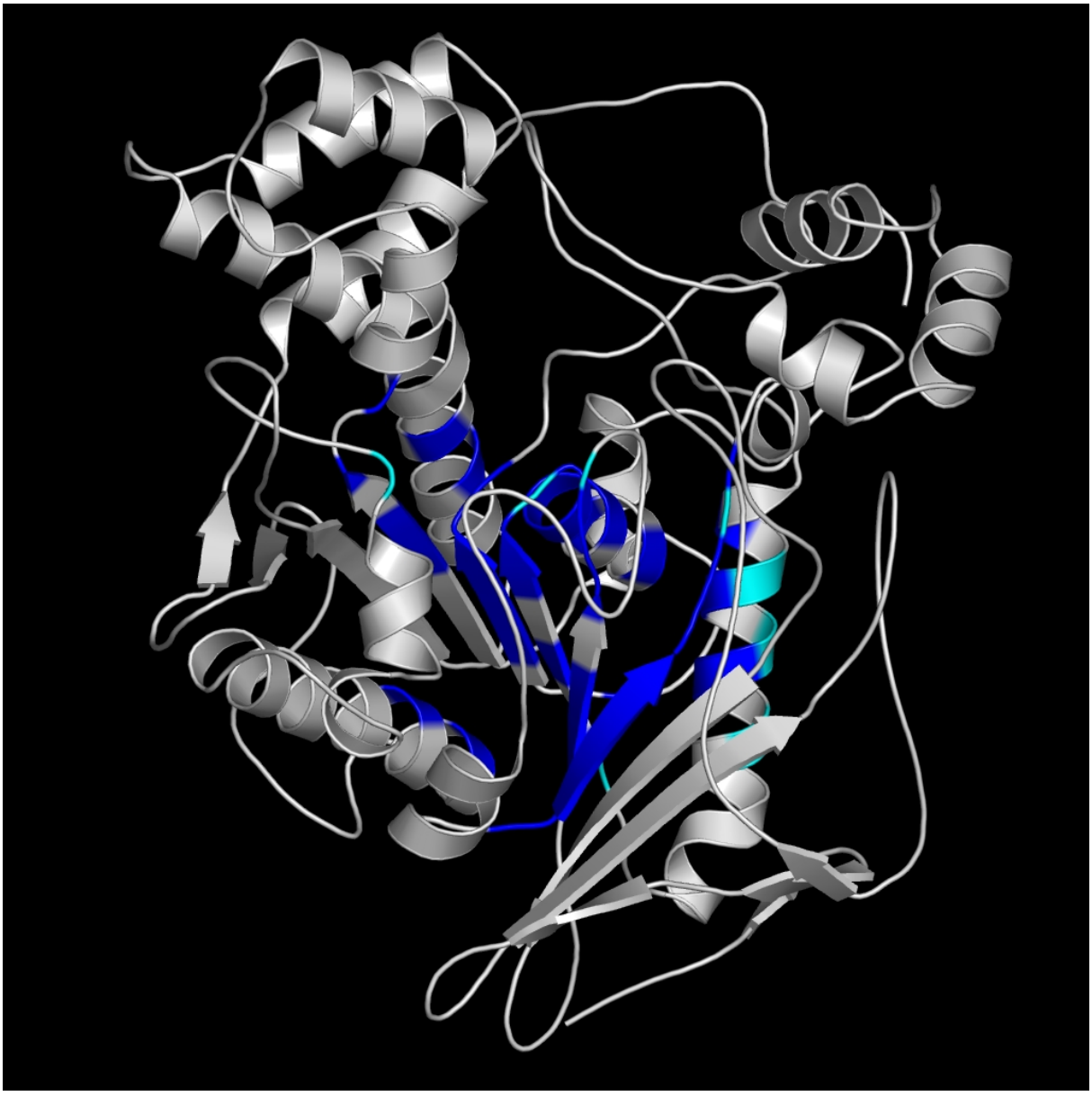


Abbildung 27: Abbildung aller 54 SKPs auf die Referenzstruktur 1AKN der Bile-salt aktivierten Lipase aus *Bos taurus*. Für das Lösungsmittel unzugängliche Positionen sind blau dargestellt, zugängliche Positionen cyan.

In Abbildung 27 wurden alle 54 SKPs auf der Struktur 1AKN abgebildet. Die für das Lösungsmittel zugänglichen SKPs wurden in cyan, die restlichen SKPs in blau dargestellt. Sämtliche SKPs waren im *Core*-Bereich der α/β-Hydrolasen zu finden wobei 72% der SKPs

in Bereichen mit definierter Sekundärstruktur lagen. Neben den mittleren β-Strands des zentralen β-Faltblatts waren vor allem die inneren Aminosäuren der amphiphilen Crossover Helices konserviert. Bildet man die SKPs auf die schematische Darstellung der α/β-Hydrolase Domäne der 1AKN ab zeigt sich, dass die Konservierung überwiegend in den Sekundärstrukturelementen β3, β4, β5, β6, αB und αC lokalisiert war (Abbildung 28). Dabei zeigte sich, dass die Konservierung innerhalb der β-Strands oszillierte, so dass überwiegend die Aminosäuren der β-Strands β3, β4 und β5 streng konserviert waren, die mit den auf einer gemeinsamen Seite des zentralen β-Faltblatts (hinter der Papierebene in Abbildung 28) liegenden α-Helices αB und αC wechselwirkten.

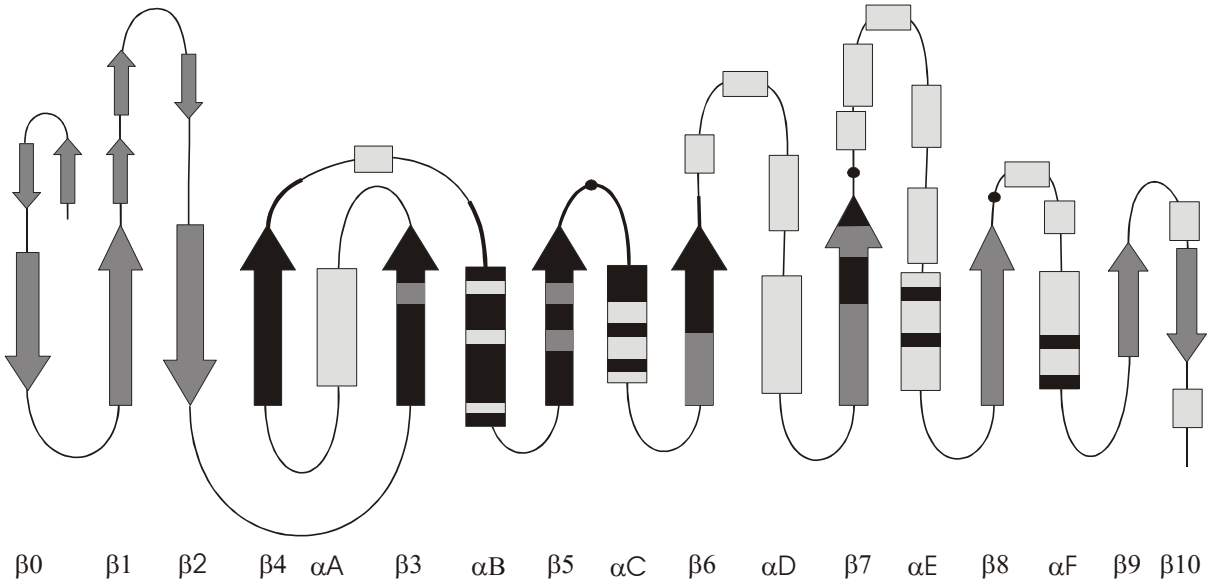


Abbildung 28: Abbildung aller 54 SKPs (schwarz) auf die topologische Darstellung des α/β-Hydrolase Folds der Referenzstruktur 1AKN der Bile-salt aktivierten Lipase aus *Bos taurus*. α-Helices sind als Rechtecke dargestellt, β-Strands als Pfeile, die Aminosäuren der katalytischen Triade als Punkte.

Hierbei waren folgende SKPs involviert: β3 (SKP_1, SKP_3, SKP_4), β4 (SKP_9, SKP_{11}, SKP_{13}), β5 (SKP_{26}, SKP_{27}, SKP_{28}), αB (SKP_{18}, SKP_{19}, SKP_{20}, SKP_{23}, SKP_{24}) und αC (SKP_{35}, SKP_{37}). Mit Ausnahme von SKP_9, SKP_{19} und SKP_{24} waren alle weiteren SKPs mit hydrophoben Aminosäuren besetzt, wobei sich zeigte, dass das SKP_{23} entweder von den Aminosäuren Isoleucin, Leucin, Valin oder Methionin dominiert war oder mit sterisch anspruchsvollen Aminosäuren wie Phenylalanin, Tryptophan oder Tyrosin besetz war. Die mit diesen SKPs besetzten Sekundärstrukturelemente stellen somit den eigentlichen hydrophoben Kern der α/β-Hydrolasen dar, der in Abbildung 29 im Detail dargestellt ist. Hier wird die einseitige Konservierung der Aminosäuren in den zentralen β-Strands deutlich mit Ausnahme des

β-Strands β4 der durchgängig streng konservierte Positionen enthält. Die den α-Helices αB und αC zugewandten SKPs der β-Strands bilden eine hydrophobe Fläche, die mit den zu dieser Fläche quer stehenden Seitenketten der α-Helix SKPs wechselwirkt. Neben diesen hydrophoben Wechselwirkungen wird diese Supersekundärstruktur auch über hydrophile Wechselwirkungen stabilisiert. Die in der α-Helix αB lokalisierte und dem Lösungsmittel zugängliche SKP_{19} (siehe oben) interagiert über hydrophile Wechselwirkungen mit der C-terminal des β-Strands β4 lokalisierten SKP_{16}. In 5 Superfamilien (abH1, abH4, abH9, abH14, abH27) konnte die Ausbildung einer Salzbrücke beobachtet werden während in abH15 eine Wasserstoffbrücke zur Stabilisierung beiträgt.

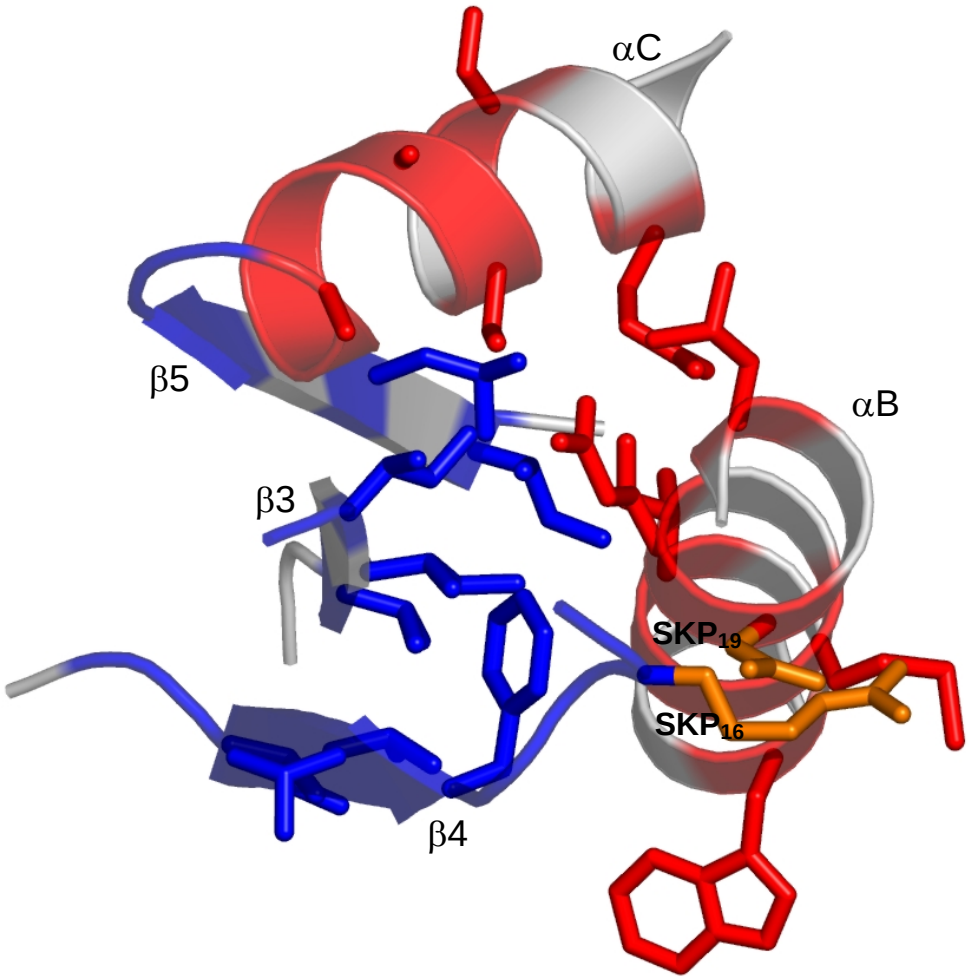


Abbildung 29: Darstellung des hydrophoben Kerns (β-Strands β3, β4, β5 und α-Helices αB, αC) des α/β-Hydrolase Folds am Beispiel der Referenzstruktur 1AKN. Die SKPs sind farbig auf die Sekundärstrukturen abgebildet, blau β-Strands, rot α-Helices. Die Interaktion zwischen SKP_{16} und SKP_{19} ist in orange hervorgehoben.

3.5.5 Die Anatomie der Fettsäurebindungsstelle von Lipasen

Zur Analyse der Fettsäurebindungsstelle der Lipasen wurden Referenzstrukturen für die homologen Familien abH1.4, abH3.1, abH8.9, abH15.2, abH20.3, abH23.1, abH36.3 und abH37.1 gewählt (Tabelle 5).

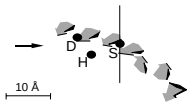
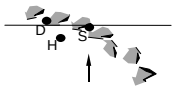
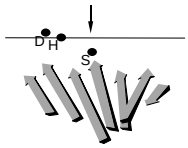
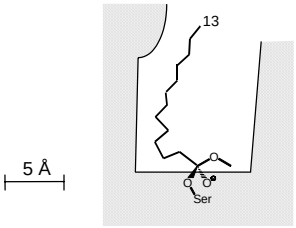
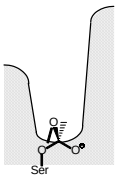
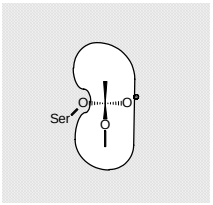
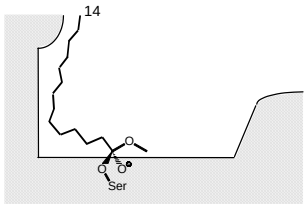
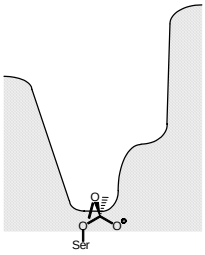
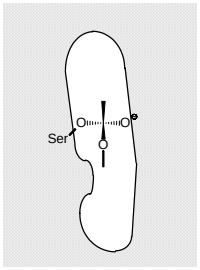
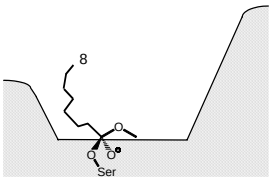
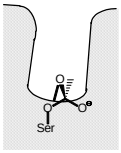
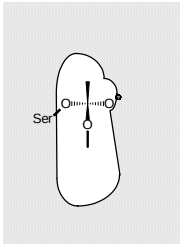
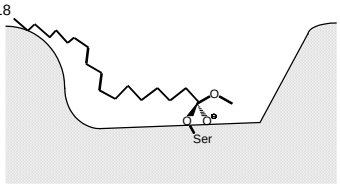
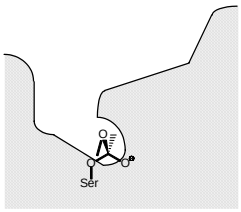
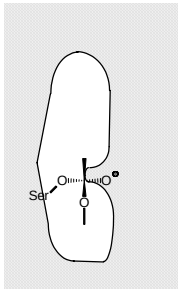
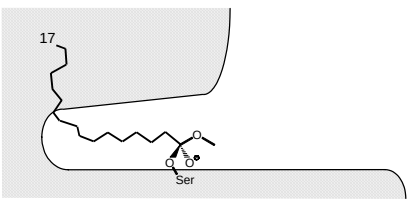
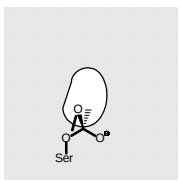
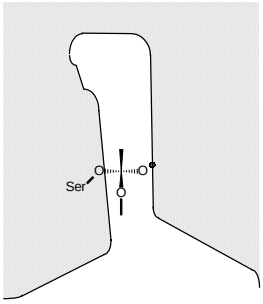
Tabelle 5: Liste der repräsentativen α/β-Hydrolasen, deren Fettsäurebindungsstelle untersucht wurde

Enzym	PDB Eintrag	homologe Familie	Substrat	katalytische Triade	*Oxyanion Hole*	Ref.
Candida antarctica Lipase	1LBS	abH37.1	Triglyceride	S105, H224, D187	T40, Q106	[100]
Burkholderia cepacia Lipase	2LIP	abH15.2	Triglyceride	S87, H286, D264	L17, Q88	[101]
Humane Pankreaslipase	1LPB	abH20.3	Triglyceride	S153, H264, D177	F78, L154	[102]
Rhizomucor miehei Lipase	5TGL	abH23.1	Triglyceride	S144, H257, D203	S82, L145	[103]
Candida rugosa Lipase	1LPO	abH3.1	Triglyceride	S209, H449, E341	G124, A210	[104]
Fusarium solani Cutinase	1XZM	abH36.3	Cutin, Triglyceride	S120, H188, D175	S42, Q121	[105]
Torpedo californica Acetylcholinesterase	1AMN	abH1.4	Acetylcholin	S200, E327, H440	G119, A201	[106]
Streptomyces aureofaciens Bromoperoxidase	1BRO	abH8.9	Acetate	S98, H257, D228	F32, M99	[107]

In allen untersuchten α/β-Hydrolasen ist die Bindungsstelle in einer Tasche über dem zentralen β-Faltblatt lokalisiert. Für Lipasen ist die Proteinoberfläche an den Rändern dieser Tasche hydrophob. Diese Regionen sollen mit der hydrophoben Substratgrenzfläche wechselwirken. Esterasen und Lipasen unterscheiden sich in der Größe dieser hydrophoben Bereiche sowie in der Form, Tiefe und den physikalisch-chemischen Eigenschaften der Substratbindungstasche. Zur Charakterisierung der Substratbindungstaschen dieser hier untersuchten Lipasen wurden Fettsäurelipasekomplexe modelliert, wobei Teile der Fettsäuren nicht berücksichtigt wurden, wenn sie außerhalb der Bindungstasche zu liegen kamen.

Die Substratbindungstasche der Lipase B aus *Candida antarctica* (CALB) wird durch einen elliptischen und steil abfallenden Trichter mit den Abmessungen 9.5 x 4.5 Å beschrieben. Das

Substrat ist parallel zur Längsachse ausgerichtet. Betrachtet man den Bindungskanal entlang der Alkohol-Fettsäureachse zeigt sich, dass die begrenzenden Wände sich in der Höhe unterscheiden (Abbildung 30b, Frontansicht). Die Höhe der linken Wand beträgt 6 Å. Das Lid bildet die rechte Wand (10,5 Å). Der experimentell bestimmte Phosphonatinhibitor der CALB kommt mit seinem Ethanolrest in dem als METHYL Bindungsstelle vorhergesagten Bereich der Fettsäurebindungsstelle zu liegen (Abbildung 31b). Somit lässt sich bis zum C4 der Fettsäure deren Lage anhand des Inhibitors ableiten. Sie binden in einen engen Spalt (wie auch von der METHYL Sonde vorhergesagt) am hydrophilen Grund des Trichters, der durch D134 und dem katalytisch aktiven S105 gebildet wird. Die linke Seite dieses Spaltes wird durch T138, I189 und V190 gebildet, während auf der rechten Seite Q157 und das *oxyanion hole* bildende T40 liegen. Am Ende des Spaltes, nahe Q157, knickt die Fettsäure stark ab und folgt dem Verlauf der linken Wand (Abbildung 30b, Seitenansicht). Bis zum C7 der Fettsäure wird die Bindungsstelle durch einen engen Spalt geformt. C7 bis C13 sind in einem hydrophoben Bereich lokalisiert, der durch V154, I285, L144 und V149 gebildet wird. Mit zunehmenden Abstand zum aktiven Zentrum verjüngt sich die Bindungstasche und findet ihr Minimum am C13 der Fettsäurekette. Hier, 13,5 Å über dem Grund, ist das Ende des Trichters erreicht und öffnet sich in die hydrophobe Oberfläche des Enzyms, die vermutlich mit der hydrophoben Substratgrenzfläche in Kontakt steht. Die Höhe der Wand der Alkoholbindungsstelle beträgt 10,5 Å. Das Substrat gelangt von oben rechts der Frontansicht (Abbildung 30b, Frontansicht) in die Bindungstasche.

Ansichten a	 Seitenansicht	 Frontansicht	 Aufsicht
Candida antarctica Lipase B b			
Burkholderia cepacia Lipase c			
Humane Pankreaslipase d			
Rhizomucor miehei Lipase e			
Candida rugosa Lipase f			

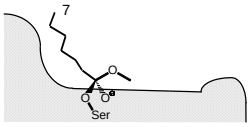
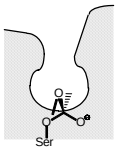
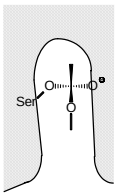
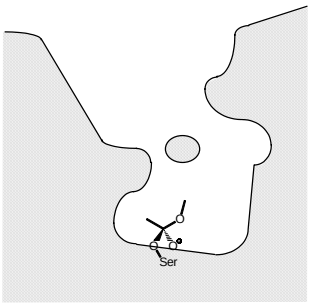
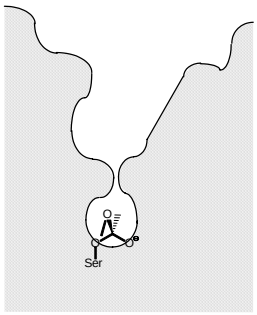
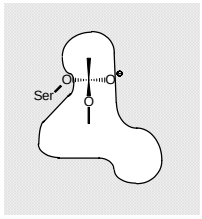
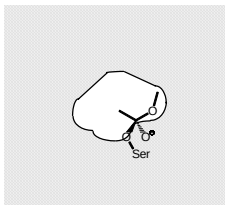
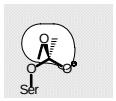
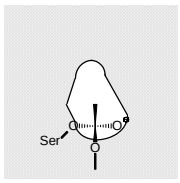
Fusarium solani Cutinase g			
Torpedo californica Acetylcholin esterase h			
Streptomyces aureofaciens Bromoperoxidase i			

Abbildung 30: Form der Bindungstasche von acht α/β-Hydrolasen. (a) Orientierung der Querschnitte, die Ebenen senkrecht zur Papierebene darstellen und durch eine gerade Linie angedeutet werden. Die Sichtachse ist durch einen Pfeil dargestellt. (b)-(i) Form der Bindungstasche in der Seiten-, Front- und Aufsicht. Für (b)-(g) ist ein Modell einer Fettsäurekette dargestellt, der Alkoholrest des eigentlichen Estersubstrats ist der Übersicht wegen nicht dargestellt. Die Zahlen geben die Länge der längsten Fettsäurekette an, die vollständig in der Tasche binden.

Die Bindungstasche der Lipase aus *Burkholderia cepacia* (BCL) stellt einen elliptischen Trichter dar mit einer Länge von 17 Å. Am Grund des Trichters entspricht die Breite mit 4,5 Å der der CALB. Der Trichter erweitert sich bis zum Eingang der Bindungstasche auf 10,5 Å. Die Höhe der linken und rechten Wand der Bindungstasche entlang der Alkohol-Fettsäure Achse beträgt 10,5 Å bzw. 16,5 Å (Abbildung 30c, Frontansicht). Das Lid wird durch die rechte Wand gebildet. Das Substrat bindet bis zum C6 der Fettsäure in einem engen Spalt (Abbildung 31c) am hydrophoben Grund des Trichters, der durch P113 und dem katalytischen S87 gebildet wird. Der Spalt wird durch V266 und V267 auf der linken und durch L167 und L17 der *oxyanion hole* Aminosäure auf der rechten Seite gebildet (Abbildung 30c, Frontansicht). Am Ende des Spaltes, nahe S117 und A120, knickt die Fettsäure stark ab und bindet entlang der linken Wand (Abbildung 30c, Seitenansicht) in einer weiteren Spalte bis zum C9 der Fettsäure und von C10 bis C14 in einem hydrophoben Bereich, der durch F119,

V123 und L164 gebildet wird. Im Gegensatz zur CALB verjüngt sich dieser hydrophobe Bereich nicht. In 10,5 Å Höhe endet der Trichter der Bindungstasche, die einem Überhang folgend die Proteinoberfläche erreicht. Die Alkoholbindungsstelle ist niedriger und steigt flach bis auf 5 Å an. Das Substrat gelangt in der Frontansicht von oben (Abbildung 30b, Frontansicht) in die Bindungstasche.

Die trichterförmige Bindungstasche der humanen Pankreas-Lipase (HuPL) (13 x 4,5 Å am Grunde der Bindungstasche) ist flacher (7 Å) als die der BCL. Ihre linke Wand (Abbildung 30d, Seitenansicht) ist die niedrigste aller untersuchten Lipasen (4,5 Å). Des Weiteren ist HuPL die einzige Lipase in der die linke Wand der Seitenansicht niedriger ist, als die linke und rechte Wand der Frontansicht. Das Lid wird im Unterschied zu den anderen Lipasen durch die linke Wand der Alkoholbindungsstelle gebildet (Abbildung 30d, Frontansicht) und nicht durch die rechte Wand. Der β5-Loop liegt gegenüber dem Lid and stellt die rechte Wand der Alkoholbindungsstelle dar (Abbildung 30d, Frontansicht). Die Position des Alkylrestes des experimentell bestimmten Undecanphosphonatinhibitors wurde zur Modellierung einer C8 Fettsäurekette genutzt. C2 bis C4 der Fettsäurekette bindet in einem Spalt am Grunde des Trichters anhand der vorhergesagten METHYL Sonde Bindungsstelle (Abbildung 31d). Dieser Spalt wird durch A179 und dem katalytischen S153 gebildet. Am Ende des Spaltes knickt die Fettsäurekette ab und folgt einer hydrophoben Bindungsstelle beginnend mit C6 entlang der linken Wand (Abbildung 30d, Seitenansicht), die durch P181 gebildet wird und endet bei C7. Sowohl die linke als auch die rechte Wand sind hydrophob (Abbildung 30d, Frontansicht). C8 der Fettsäurekette bindet an die rechte Wand der Frontansicht, die durch Q157 und die *oxyanion hole* Aminosäure F78 am Ende des Trichters gebildet wird. Die linke Wand wird durch F216 und I210 gebildet. Die Alkoholbindungsstelle ist mit 10,5 Å höher als in BCL. Das Substrat gelangt in der Frontansicht von oben (Abbildung 30d, Frontansicht) in die Bindungstasche.

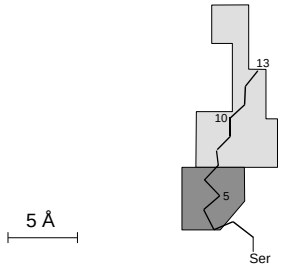
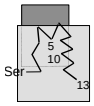
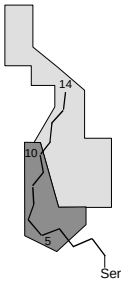
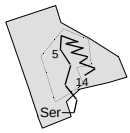
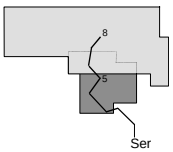
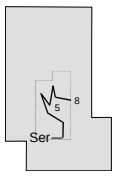
Die Bindungstasche der Lipase aus *Rhizomucor miehei* (RML) gleicht einer flachen Wanne mit 18 Å Länge und einer Breite von 4,5 Å am Grund der Bindungstasche, die sich an der Proteinoberfläche auf 6 Å weitet. Die linke Wand (Abbildung 30e, Frontansicht) ist 8 Å hoch. Die rechte Wand, die das Lid darstellt, ist 12 Å hoch. Der Alkylrest des experimentell bestimmten Hexylphosphonatinhibitors bindet in die durch die METHYL Sonde vorhergesagte Bindungsstelle. Die Fettsäurekette bindet bis zu C8 in einer Spalte am Grund der Bindungstasche (Abbildung 31e). P177, H108 und das katalytische S144 bilden den Grund, während V205 und D91 links bzw. rechts liegen. C8 bis C10 der Fettsäurekette binden entlang der rechte Wand (Abbildung 30e, Seitenansicht) und folgt einem hydrophoben Spalt

bis zum C18 der Fettsäurekette (Länge beträgt 9,5 Å), der parallel zur Proteinoberfläche verläuft. Seine Breite von 5,5 Å am Grund gibt dem Substrat Bewegungsfreiheit. Dieser wird durch die Aminosäuren P209 und P210 gebildet, seine linke und rechte Wand durch L208 bzw. F94/F213. Die Höhe der Alkoholbindungsstelle beträgt 8 Å. Das Substrat gelangt in der Frontansicht von oben in die Bindungstasche.

Die Bindungstasche der Lipase aus *Candida rugosa* (CRL) ist außergewöhnlich und unterscheidet sich von denen der anderen Lipasen deutlich. In der CRL bindet die Fettsäurekette in einem Tunnel der im Inneren des Proteins liegt und der sich in einen weiten Zugang öffnet (Abbildung 30f, Seitenansicht). Dieser Tunnel ist mindestens 22 Å lang mit einem Durchmesser von ca. 4 Å. In der unkomplexierten CRL ist die Bestimmung des Tunnels schwierig, da dieser durch Seitenketten blockiert ist. Der Tunnel wird gebildet durch G124, (*oxyanion hole*), F125, dem katalytischen S209, A210 (*oxyanion hole*), M213, V245, P246, F296, S301, L302, R303, L304, L307, F345, Y361, F362, S365, F366, V409, L410, L413, G414, F415, F532 und V534. Die Länge des Hexadecylrestes des experimentell bestimmten Inhibitors entspricht einer C17 Fettsäurekette. Das katalytisch aktive Serin liegt kurz hinter dem Eingang zum Tunnel, der sich bei C3 verjüngt und die Fettsäurekette bis zum ω-Ende eng bindet. Die hydrophobsten Bereiche des Tunnels befinden sich in den Regionen um C4 bis C7 und C12 bis C14. Im Unterschied zu anderen Lipasen ist die Alkoholbindungsstelle nicht durch eine rechte Wand begrenzt (Abbildung 30f, Seitenansicht). Das Lid liegt auf der unteren, rechten Seite (Abbildung 30f, Aufsicht) wie in anderen Lipasen auch. Das Substrat gelangt von rechts in die Bindungstasche (Abbildung 30f, Seitenansicht).

Die Substratbindungstasche der Cutinase aus *Fusarium solani* hat eine Länge von 16 Å und eine Breite von 4,5 Å am Grund der Bindungstasche und verjüngt sich auf 2 Å an der Proteinoberfläche. Die linke und rechte Wand sind 6 bzw. 8 Å hoch. Eine Fettsäurekette mit einer Länge von C7 wurde anhand des experimentell bestimmten Undecanphosphonat-inhibitors in die Bindungstasche modelliert. Bis C3 bindet diese in einem Spalt am Grund der Bindungstasche (METHYL Sonde), der durch T150, dem katalytischen S120 und dem katalytischen H188 gebildet wird. Die linke Seite wird durch V177/V184 und die rechte Seite durch N84/S42 (*oxyanion hole*) gebildet (Abbildung 30g, Frontansicht). Die Fettsäurekette folgt der linken Wand (Abbildung 30g, Seitenansicht), die durch L182 gebildet wird. Hier, am C7 der Fettsäurekette endet die trichterförmige Bindungstasche. Die Bindungstasche ist mäßig hydrophob. Die hydrophobsten Bereiche liegen an der Proteinoberfläche, die vermutlich mit der Substratgrenzfläche wechselwirken. Die Alkoholbindungsstelle ist sehr flach.

Trotz der hohen Sequenzähnlichkeit zwischen der Acetylcholinesterase aus *Torpedo californica* (AChE) und CRL, unterscheiden sich die Bindungstaschen dieser Enzyme deutlich. Die Substratbindungsstelle der AChE ist tief im Inneren des Enzyms lokalisiert (Abbildung 30h, Seiten- und Frontansicht) mit einer kleinen Acyl- und einer großen Alkoholbindungsstelle die 3,5 bzw. 7,5 Å breit sind (Abbildung 30h, Aufsicht). Gemessen entlang der Alkohol-Säureachse beträgt die Länge 10 Å. Das Substrat gelangt durch einen engen Kanal von unten rechts (Abbildung 30h, Aufsicht) in die Bindungstasche.
Die Bindungstasche der Bromoperoxidase aus *Streptomyces aureofaciens* ist ein kleiner, elliptischer Hohlraum im Inneren des Proteins. Dieser Hohlraum ist 7 Å lang (entlang der Alkohol-Säureachse), 4 Å breit und 6,5 Å hoch. Die Alkoholbindungsstelle ist sehr klein. Die Säurebindungsstelle ist kurz, knickt nach oben links ab (Abbildung 30i, Frontansicht) und öffnet sich in einen engen Kanal der zur Proteinoberfläche führt. Jedoch mögen diese Abschätzungen anhand des freien Enzyms die Abmessungen unterbewerten.

Ansichten a	Seitenansicht	Aufsicht
Candida antarctica lipase B b		
Burkholderia cepacia Lipase c		
Human pancreatic lipase d		

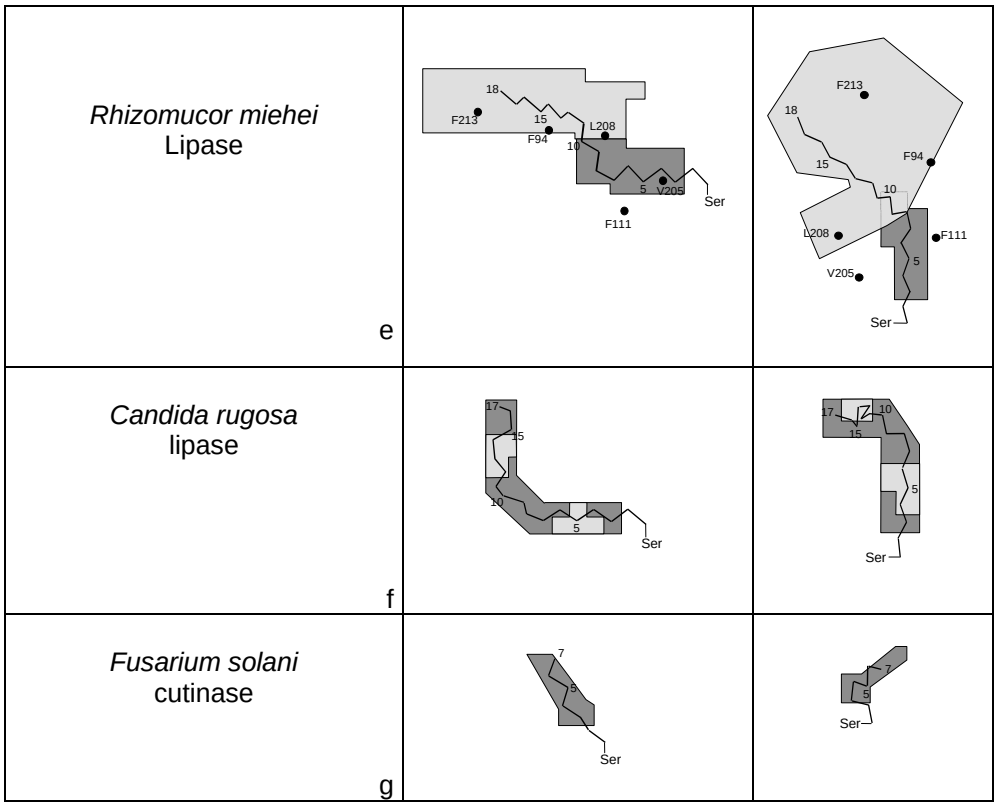


Abbildung 31: Schematische Darstellung der GRID Potentiale (Methyl Sonde: dunkelgrau, DRY Sonde: hellgrau) sowie der Position der Fettsäurekettenmodelle. Die Ansichten entsprechen denen in Abbildung 30. Für RML (e) sind Aminosäurepositionen angegeben, die für die Kettenlängenspezifität der homologen RDL und ROL verantwortlich sind.

3.5.6 Die Architektur des *Oxyanion Hole*: drei Hydrolase Klassen

Das *oxyanion hole* der Lipasen besteht aus zwei Aminosäuren, die über ihre Backboneamidwasserstoffe oder die Hydroxylgruppe einer Seitenkette das Substrat in seinem Übergangszustand stabilisieren. Eine Aminosäure befindet sich im strukturell konservierten *nucleophilic elbow* und folgt direkt dem katalytisch aktiven Nucleophil. Daraus ergibt sich die Konsequenz, dass die Orientierung der Backboneamidbindung dieser Aminosäure in allen Lipasen identisch ist. Im Gegensatz dazu befindet sich die weitere *oxyanion hole* bildende Aminosäure in einem Loop nach dem β-Strand β3, der weder in der Sequenz noch in der Struktur konserviert ist. Die Analyse der Superfamilien, für die eine Proteinstruktur verfügbar war und somit diese *oxyanion hole* bildende Aminosäure identifiziert werden konnte, zeigte jedoch, dass für 20 Superfamilien (abH8, abH9, abH12, abH13, abH14, abH15, abH18, abH19, abH20, abH22, abH23, abH25, abH26, abH31, abH32, abH33, abH34, abH35, abH36, abH37) der Loop strukturell ähnlich war, und die Position der *oxyanion hole* bildenden Aminosäure streng erhalten war. Dieser Aminosäure ging immer ein streng konserviertes

Glycin voraus. Diese *oxyanion hole* Architektur erhielt deshalb die Bezeichnung GX Typ, wobei X die *oxyanion hole* Aminosäure darstellt. Die Superfamilien, die diesen GX Typ des *oxyanion holes* aufwiesen, wurden der GX Klasse zugeordnet. Die Familie abH31 stellt eine Ausnahme dar, da hier das Glycin gegen ein streng konserviertes Glutamat ausgetauscht war. Dieses Glutamat wechselwirkte mit einem 5,1 Å entfernten Arginin, was zur Ausbildung einer Salzbrücke führen kann. Das Arginin war innerhalb der homologen Familie abH31.2 mit bekannter Proteinstruktur streng konserviert. Die Analyse des Multisequenz Alignments der Superfamilie zeigte, dass das Arginin nicht streng konserviert war, jedoch waren für die weiteren homologen Familien in naher Nachbarschaft zu diesem Arginin ebenfalls streng konservierte Arginine zu finden. Die Familie abH31 wurde Aufgrund der strukturellen Ähnlichkeit des *oxyanion hole* bildenden Loops sowie dessen Stabilisierung (siehe folgender Absatz) ebenfalls der GX Klasse zugeordnet. Aus den Verwandtschaftsbeziehungen, die über den Vergleich der HMMs abgeleitet wurden (Absatz 3.5), konnten 7 weitere Superfamilien der GX Klasse zugewiesen werden (abH7, abH10, abH11, abH16, abH17, abH21, abH24). Diese Superfamilien enthielten zwar keine Proteinstrukturdaten, wiesen jedoch signifikante Sequenzähnlichkeit im Bereich des *oxyanion holes* zu den Superfamilien der GX Klasse auf.
Während die Orientierung der Backboneamidbindung in allen offenen Lipasestrukturen erhalten war, unterschied sich die Aminosäure X in ihren physikalisch-chemischen Eigenschaften und der Orientierung der Seitenkette. Die Analyse der Multisequenz Alignments für diese Superfamilien und den zugehörigen homologen Familien ergab, dass die physikalisch-chemischen Eigenschaften der Aminosäure X innerhalb der Familien erhalten waren: hydrophil in abH8.[5,10] (E, N), abH12.3 (E), abH16 (T), abH22.7 (S), abH23 (T, S), abH36 (T, S, R), abH37 (T) und hydrophob in abH7 (L, F), abH8.[1,2,3,4,6,7,9,13] (F, W), abH9 (W), abH10 (I), abH11 (F), abH12.[1,2] (A, I, L), abH13 (A), abH15 (F, L), abH17 (Y), abH18 (I), abH19 (I,M,L), abH20 (W, F), abH21 (L, V, Y, W), abH22.[1,2,3,4,5,6,8,9] (L, W), abH25 (F), abH26 (Y), abH28 (Y), abH32 (I), abH33 (L), abH35 (F) sowie ein Glycin in abH8.8, abH9.3, abH32.2, abH34. Ausnahmen stellten die Familien abH8.11 (S, A) und abH8.14 (K, I) dar, für die innerhalb der Familie sowohl hydrophile als auch hydrophobe Aminosäuren an der Position X festgestellt wurden. Für die Familie abH8.12 konnte trotz signifikanter Sequenzähnlichkeit zur Familie abH8 kein typische GX Motiv identifiziert werden. Diese Familie enthält jedoch nur hypothetische Proteine deren enzymatischen Eigenschaften noch nicht experimentell bestimmt wurden.

Die Superfamilien abH1, abH3, abH4 und abH6 unterscheiden sich von den Lipasen der GX Klasse. Strukturüberlagerungen ergaben, dass die *oxyanion hole* bildende Aminosäure im Vergleich zur GX Klasse um eine Position zum C-Terminus hin verschoben ist (Abbildung 32) und durch ein Glycin oder Alanin gebildet wird. Dieser Aminosäure folgt eine konservierte hydrophobe Aminosäure X. Diese Architektur des *oxyanion holes* wurde als GGGX Typ bezeichnet und die Superfamilien, die dieses Motiv trugen in der GGGX Klasse zusammengefasst. Über die aus den HMM Vergleichen abgeleiteten Verwandtschafts-beziehungen zwischen den Superfamilien konnten auch die Superfamilien abH2 und abH5 der GGGX Klasse zugeordnet werden. Diese wiesen eine signifikante Sequenzähnlichkeit im Bereich des *oxyanion holes* zu den Superfamilien der GGGX Klasse auf. Die Analyse der Multisequenz Alignments der GGGX Familien zeigte, dass für 11% (51 Proteine) der GGGX Proteineinträge das *oxyanion hole* bildende Glycin oder Alanin durch andere Aminosäuren ersetzt war. Davon waren 43% putative Proteineinträge mit überwiegend hydrophilen Variationen für den *oxyanion hole* Bildner (S, R, E) und 14% nicht katalytisch aktive Neuroligine und Gliotactin (S, E). Die verbleibenden Proteineinträge wiesen ebenfalls überwiegend hydrophile Aminosäuren an dieser Position auf (S, T, R, D), vereinzelt aber auch hydrophile Aminosäuren (V, C, I).

In den meisten Fällen entspricht die hydrophobe Aminosäure X einem Phenylalanin, Leucin, Tyrosin, Tryptophan oder Methionin. Jedoch kommen auch hier Abweichungen vor wie in der homologen Familie Yarrowia lipolytica in der für 2 Proteineinträge X durch ein Asn gebildet wird. Wenn sich auch die Struktur des *oxyanion hole* tragenden Loops für GX und GGGX Hydrolasen unterscheiden, ist die Position des Wasserstoffatoms der Backboneamidgruppe, das für die Stabilisierung des Übergangszustandes verantwortlich ist, erhalten, und somit auch die Funktion (Abbildung 32).

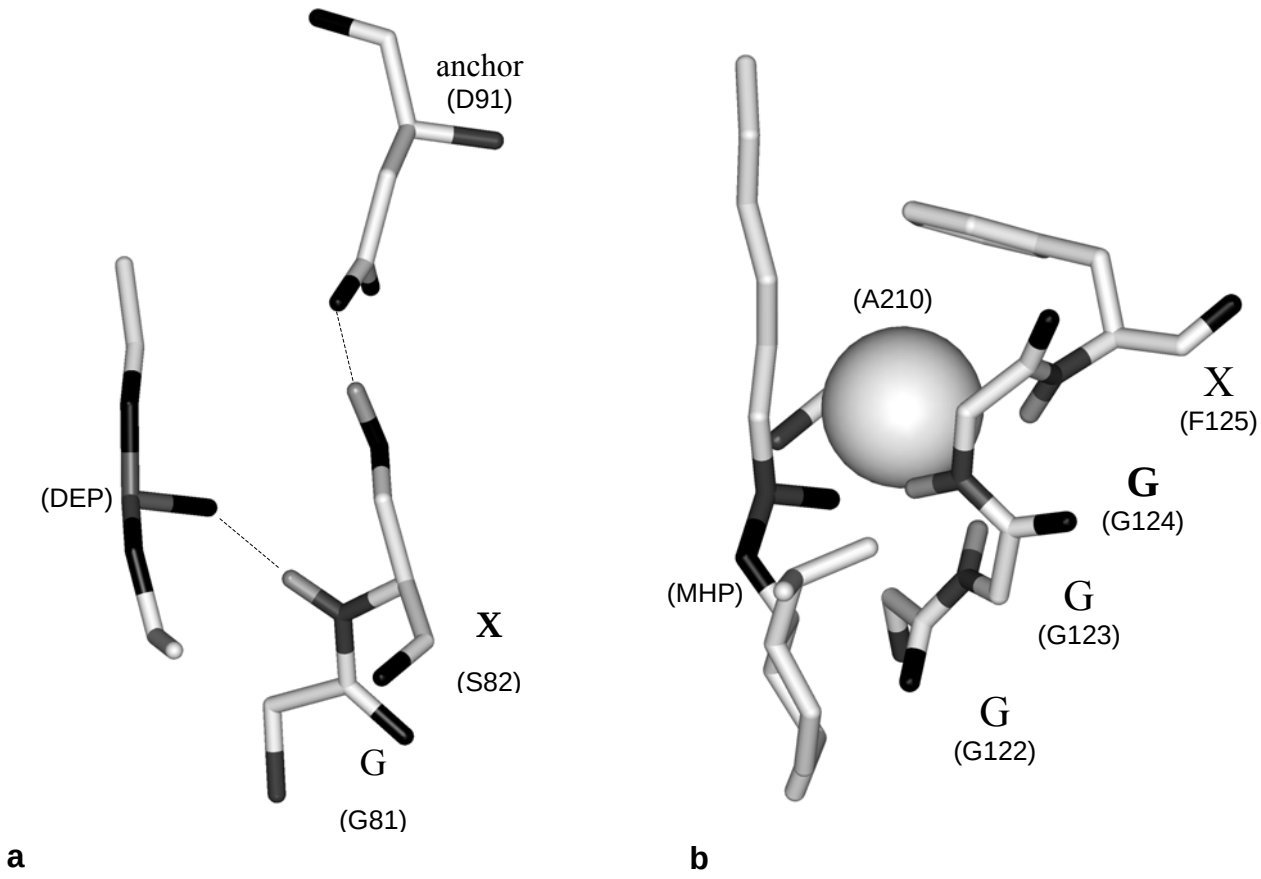


Abbildung 32: Darstellung der Architektur des *oxyanion holes* der GX-Klasse und GGGX-Klasse. (a) GX-Typ für RML (4TGL): Stabilisierung des substratanalogen Inhibitors Diethylphosphonat (DEP) über eine Wasserstoffbrücke mit der ersten *oxyanion hole* Aminosäure X (S82). Die Aminosäure S82 wird über eine Wasserstoffbrücke mit dem Anker D91 stabilisiert. (b) GGGX-Typ für CRL (1LPM): Stabilisierung des substratanalogen Inhibitors (1*R*)-Menthylhexylphosphonat über die erste *oxyanion hole* Aminosäure G124. Die Aminosäure G124 wird stabilisiert über die Fixierung der Seitenkette von A210 zwischen G124 und der Seitenkette der Aminosäure X.

Für die Superfamilien abH27, abH28, abH29 und abH30 ergaben Strukturuntersuchungen, dass das *oxyanion hole* nicht über ein Backboneamidwasserstoff stabilisiert wird, sondern durch die Hydroxylgruppe der Seitenkette eines Tyrosins. Dieses Tyrosin war innerhalb der Superfamilien streng konserviert, konnte aber strukturell an zwei verschiedenen Positionen lokalisiert sein. In abH27 und abH28 war das Tyrosin im direkten Anschluss an den β-Strand β3 platziert, während es in abH29 und abH30 weiter N-terminal zu β3 positioniert war. Die Analyse der Multisequenz Alignments für die Familien abH29 und abH30 ergab, dass C-terminal zum *oxyanion hole* bildenden Tyrosin ein streng konserviertes Prolin lokalisiert war.

In Abbildung 33 ist der Vergleich der beiden Tyrosinpositionen in den Familien abH28 und abH30 sowie der Familie abH26, die ein typischer Vertreter der GX-Klasse, dargestellt. Die Familie abH26 wurde zum Vergleich herangezogen, da die Position X durch ein Tyrosin besetzt war und eine nahe Verwandtschaft zu den Familien abH27 und abH28 besteht. Die Lage der Hydroxylwasserstoffatome der Oxyanion stabilisierenden Tyrosinseitenketten entsprach hierbei der des Backboneamidwasserstoffs der GX Familie. Der Verlauf des Loops zwischen dem β-Strand β3 und der α-Helix αA unterschieden sich jedoch deutlich. Um der Tyrosinseitenkette den nötigen Raum zu verschaffen die Hydroxylgruppe optimal zu positionieren, ist der β-Strand β3 sowie der sich anschließende Loop deutlich vom aktiven Zentrum entfernt.

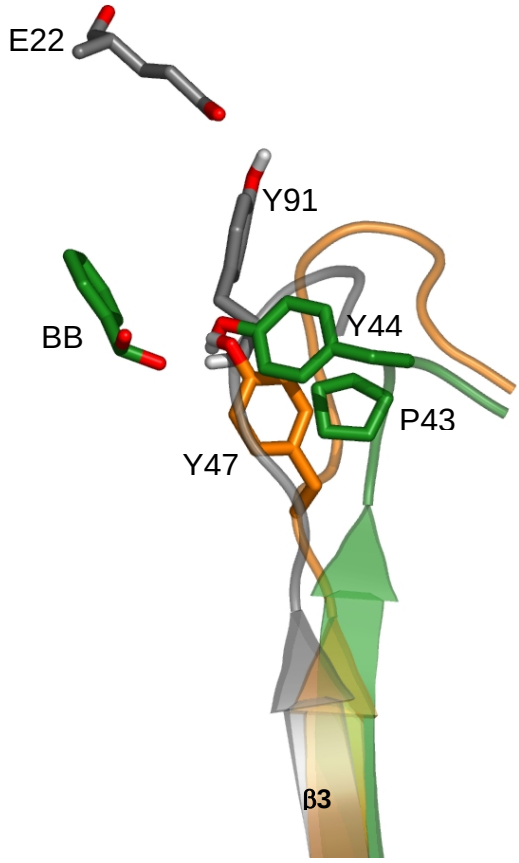


Abbildung 33: Darstellung der *oxyanion hole* Architektur der Y-Klasse (grün: Kokainesterase aus *Rhodococcus sp.* 1JU3, orange: Prolyl Endopeptidase aus *Sus scrofa* 1E8M) im Vergleich zur GX-Klasse (grau: Deacetylase aus *Bacillus subtilis* 1ODT). Für Y-Klasse erfolgt Stabilisierung des tetrahedrale Übergangszustand des Substrats (am substratanalogen Inhibitor Benzolborsäure (BBS) der Struktur 1JU3 veranschaulicht) über die Hydroxylgruppe eines Tyrosins (Y44, Y473) und nicht wie in der GX-Klasse über ein Backboneamid. Die Stabilisierung des *oxyanion hole* Tyrosins erfolgt für die Y-Klasse entweder über ein Prolin (P43) oder über die Nähe zum β-Strand β3 (Y473) und nicht über das Anker Konzept der GX-Klasse (Y91,E224).

3.5.7 Die Rolle der Seitenkette der *Oxyanion Hole* Aminosäure in der GX Klasse

Um die Rolle der Seitenkette der das *oxyanion hole* bildenden Aminosäure zu verstehen, wurden repräsentative Proteinstrukturen der GX Klasse Familien abH8.9, abH13.1, abH14.2, abH15.2, abH20.3, abH23.1, abH25.1, abH26.1, abH31.2, abH36.3 und abH37.1 untersucht (Tabelle 6). Die Strukturen mussten in der aktiven Form vorliegen, und durften an der Position X kein Glycin tragen. Obwohl die Seitenkette der Aminosäure X nicht am katalytischen Mechanismus direkt beteiligt ist, war diese innerhalb der Superfamilien sehr gut erhalten mit Ausnahme der Familie abH13.1. Der Struktureintrag 1QLW dieser Familie trägt ein Cystein als *oxyanion hole* bildende Aminosäure, während für die weiteren Familienmitglieder an dieser Position ein Alanin zu finden war

Tabelle 6: *Oxyanion hole* Aminosäure und Anker Aminosäuren der GX-Klasse. Auftretende Variationen für diese Aminosäuren innerhalb der Superfamilie sind in Klammern gesetzt.

Homologe Familie	Enzym	PDB Eintrag	*Oxyanion Hole*	Anker	Abstand (Å)	Kontakt
abH23.1	RML	4TGL	S82 (S,T)	D91 (D,N)	2,5	H-Brücke
abH15.2	BCL	2LIP	L17 (L,M,F)	L167 (L,I,V,N)	3,6	hydrophob
abH37.1	CALB	1LBS	T40	Q157 Q106 L73	3,0 3,1 3,8	H-Brücke H-Brücke hydrophob
abH36.3	FSC	2CUT	S42 (S,T,R)	N84 Q121 (N,S,D,F)	3,0 3,0	H-Brücke H-Brücke
abH8.9	SAB	1BRO	F32 (F,W)	hydrophobe Bindungstasche (P33, F61, F163, A198, A201)		hydrophob
abH25.1	SEL	1JFR	F63 (F,Y)	schwacher Kontakt zu L93		
abH13.1	ASE	1QLW	C71 (A,V)	C72	2,1	Disulfidbrücke
abH26.1	BSD	1ODS	Y91	E224	2,6	H-Brücke
abH31.2	PPC	1GGV	I37	Hydrophobe Bindungstasche (F38,Y85,W88)		Hydrophob
abH14.2	DGL	1K8Q	L67 (I,V,M)	W275 (Y,F)	4,2	Hydrophob
abH20.3	HuPL	1LPB	F78 (W,Y)	Hydrophobe Bindungstasche (Y115, A118, L154)		Hydrophob

In allen untersuchten Proteinstrukturen wechselwirkte die Seitenkette der Aminosäure X mit mindestens einer weiteren Seitenkette, die als Anker bezeichnet wurde. (Abbildung 32). In BCL und DGL ist der Anker hydrophob (L167, W275) und 3,6 Å bzw. 4,2 Å von der hydrophoben *oxyanion hole* Aminosäure (L17, L67) entfernt. In CALB, FSC, BSD und RML, für die der Anker hydrophil ist (T40, S42, Y91, S82) befindet sich der nächste Anker (Q157, N84, E224, D91) in einer Entfernung, die die Bildung einer Wasserstoffbrücke zulässt (Tabelle 6).

Sowohl der *oxyanion hole* Bildner als auch der Anker liegen entlang des Grunds des tiefen hydrophoben Spalts. In SAB, PPC und HuPL wird der Anker durch eine Gruppe von Aminosäuren gebildet. Die Seitenkette der *oxyanion hole* Bildner F32, I37 bzw. F78 ist durch eine hydrophobe Bindungstasche, die aus den Aminosäuren P33, F61, F163, A198 und A201 bzw. F38, Y85 und W88 bzw. Y115, A118 und L154 gebildet werden, fixiert. In ASE wird das *oxyanion hole* bildende Cystein über eine Cysteinbrücke mit dem benachbarten Cystein fixiert. Eine Ausnahme dieses Konzepts stellt SEL dar. Diese Lipase besitzt kein typisches Lid und der hydrophobe *oxyanion hole* Bildner F63 ist an der Protein Oberfläche exponiert. Es existiert keine hydrophobe Bindungstasche, die die Seitenkette fixieren würde und nur ein schwacher Kontakt mit L93 ist zu beobachten. Anstelle dessen zeigt die Seitenkette von F63 in Richtung der hydrophoben Substratgrenzfläche, die die Rolle des Ankers übernehmen könnte.

Wie auch der *oxyanion hole* Bildner ist der Anker von FSC, RML, BSD und BCL innerhalb der Superfamilien erhalten. Für die Superfamilie der Burkholderia Lipasen ist dieser hydrophob, wogegen für Rhizomucor Lipasen und Cutinasen der Anker hydrophil ist.

Für RML und BGL, die 84% Sequenzidentität mit BCL besitzt, sind Proteinstrukturen der geschlossen Form bekannt. In diesem inaktiven Zustand besteht kein Kontakt zwischen *oxyanion hole* Bildner und Anker. In der geschlossenen RML wechselwirkt die Seitenkette des *oxyanion hole* Bildners S82 über eine Wasserstoffbrücke mit dem Backboneamid der Aminosäure S84 und ist dadurch 9,3 Å vom Anker entfernt. Im Vergleich zur geöffneten Form der RML ist der Wasserstoff des Backboneamids um 0.6 Å weiter von der Substratbindungstasche entfernt. Dadurch ist die Lage des *oxyanion hole* nicht optimal für die Stabilisierung des Übergangszustandes des Substrats ausgerichtet.

BGL stellt ein Beispiel für Lipasen dar, die das *oxyanion hole* erst nach Öffnung des Lids ausbilden. In der geschlossenen Form der BGL erhöht sich der Abstand zwischen der Seitenkette des *oxyanion hole* Bildners und dem Anker auf 6 Å. Hierbei interagiert L17 mit dem in 3,6 Å entfernten L149 das im Lid lokalisiert ist. Dieser Kontakt führt zu einer Rotation

des Backboneamids des L17 wodurch der Wasserstoff nicht mehr zum *oxyanion hole* orientiert ist und 2,1 Å von der Lage des Wasserstoffatoms in der geöffneten Form entfernt ist.

Die Familien abH8.8, abH9.3 und abH34 teilten den selben strukturellen Aufbau des Loops in dem der *oxyanion hole* Bildner X lokalisiert ist. Da jedoch X durch ein Glycin besetzt ist, kann die Stabilisierung des Loops nicht über das Ankerkonzept erfolgen. Analysen der Multisequenz Alignments dieser Familien ergab, dass in den Familien abH9.3 und abH34 an der Position X+1 und in der Familie abH8.8 an der Position X+2 ein streng konserviertes Prolin lag. Dieses wiederum wurde gefolgt von einem bzw. zwei streng konservierten Glycine. Die Untersuchung bekannter Proteinstrukturen für die Familien abH9.3 und abH34.2 ergab, dass dieses GGPGG Motiv einen engen Turn bildet, die Stabilisierung sich jedoch zwischen diesen beiden Familien unterschied. In der Familie abH9.3 bildet die Carbonylgruppe des *oxyanion hole* bildenden Glycins zwei H-Brücken mit den Amidgruppen der dem Prolin folgenden Glycine aus. In der Familie abH34 sind diese beiden Glycine nicht in die Stabilisierung involviert. Hier bildete die Carbonylgruppe des *oxyanion hole* bildenden Glycins eine H-Brücke mit dem Backboneamid der C-terminalen Aminosäure aus. Des Weiteren folgte dem GGPGG Motiv ein Cystein, das in der Superfamilie streng konserviert war, und eine Disulfidbrücke mit einem C-terminal liegenden und ebenfalls streng konservierten Cystein bildete.

3.5.8 Stabilisierung des *Oxyanion Holes* für die GGGX Klasse

Die Stabilisierung des *oxyanion holes* für die GG**G**X Klasse unterschied sich von der der G**X** Klasse. Zusätzlich zum gut erhaltenen *oxyanion hole* Bildner **G** ist auch der C-terminale, hydrophobe Nachbar X streng konserviert. Der Vergleich repräsentativer Strukturen der GG**G**X Familien abH1 und abH3 zeigte, dass der C-terminale Nachbar des katalytisch aktiven Serins ein sehr gut erhaltenes Alanin ist. Dieses Alanin dient der Stabilisierung des *oxyanion holes*. Einem Schnappschloss gleich wird die Seitenkette dieses Alanins von den Backboneamidgruppen des *oxyanion hole* bildenden Glycins und dessen hydrophoben Nachbarn X sowie dessen Seitenkette umschlossen (Abbildung 32). Ein Sequenzvergleich der GGGX Superfamilien ergab, dass für 77% der Sequenzen dieses Alanin erhalten ist.

Vergleicht man die geöffneten Strukturen der CRL, BTL und TCA zeigt sich, dass das Dipeptid **G**X in einem β-Loop lokalisiert ist, der sowohl in der geöffneten als auch in der geschlossenen Form der Lipasen strukturell erhalten ist. Jedoch die Konformation des Dipeptids GG ändert sich durch die Aktivierung und somit Öffnung der Lipasen. In der geöffneten Form kann das Backboneamid des N-terminalen Nachbarn des *oxyanion hole*

bildenden **G** (G123 in CRL) zwei Konformationen einnehmen: Stabilisierung des Übergangszustandes[67,104] oder in Kontakt mit einem gebundenen Wassermolekül wodurch das Backboneamid vom *oxyanion hole* wegzeigt. In der geschlossen Form liegt das Dipeptid in einer stabilen β-Strandkonformation vor wobei der Wasserstoff der Backboneamidgruppe vom *oxyanion hole* wegzeigt.

3.5.9 Stabilisierung des *Oxyanion Holes* der Y Klasse

Die Stabilisierung des *oxyanion holes* der Y Klasse erfolgt in Abhängigkeit von der Lage des *oxyanion hole* bildenden Tyrosins innerhalb des Loops (Abbildung 33). In den Familien abH27 und abH28 war das Tyrosin am C-terminalen Ende des β-Strands β3 lokalisiert. Somit wurde die Position des Backbones über das zentrale β-Faltblatt bestimmt. Für die Familien abH29 und abH30 lag das Tyrosin im Loop, der sich dem β-Strand β3 anschließt. Diesem Tyrosin ging in beiden Superfamilien ein streng konserviertes Prolin voraus, welches einen scharfen Turn verursachte und so den Backbone des Tyrosin für eine optimale Positionierung der Seitenkette ausrichtete.

3.5.10 Das Anker Konzept

Für die GX-Klasse wurde in dieser Studie gezeigt, dass die Stabilisierung des Loops, der die *oxyanion hole* bildende Aminosäure X trägt, über das Ankerkonzept bewerkstelligt wird, mit Ausnahme der Familien, die ein Glycin für X tragen. Das Ankerkonzept beschreibt die Interaktion der Seitenkette der Aminosäure X mit mindestens einer Aminosäure, die als Anker bezeichneten wird. Diese Interaktion führt zur Fixierung der *oxyanion hole* Aminosäure und somit zur aktiven Orientierung des Backboneamids, das den Übergangszustand des Substrats stabilisiert. Um ein besseres Verständnis für das Ankerkonzept zu erlangen, wurden repräsentative Strukturen der Familien untersucht, die einen dezidierten Anker aufwiesen: RML, CALB, FSC und BSD mit hydrophilen Ankern sowie BCL und DGL mit einem hydrophoben Anker. Für RML und FSC war der Anker in einer α-Helix zwischen dem β-Strand β4 und der α-Helix αD lokalisiert. Für CALB, BSD, BCL und DGL war der Anker im Bereich zwischen dem β-Strand β6 und der α-Helix αD lokalisiert. Dabei zeigte sich, dass für die Strukturen der RML, CALB, FSC, BCL und DGL die topologisch unterschiedlichen und ankertragenden Bereiche strukturell überlagert sind. Der RMSD der Position des C_α-Atoms des Anker variiert zwischen 0,8 Å und 3,2 Å (Tabelle 7). Die Seitenketten des Ankers und der *oxyanion hole* bildenden Aminosäure überbrücken so einen Abstand zwischen deren C_α-Atome von 7,9 Å (RML) bis 10 Å (BCL). Eine Ausnahme

stellte BSD dar. Zwar war für dieses Enzym wie für CALB und BCL der Anker im selben topologischen Bereich lokalisiert, der Abstand zwischen den C_α-Atomen des Ankers und der *oxyanion hole* bildenden Aminosäure betrug jedoch 12,3 Å und die Position des Ankers war nicht strukturell überlagert. Diese größere Distanz wird durch die im Vergleich zu den anderen Anker-*oxyanion hole* Paaren längeren Seitenketten eines Tyrosins und Glutamats überbrückt.

Tabelle 7: RMSD der C_α-Atome der Anker Aminosäuren

	RML	CALB	FSC	BCL	DGL
RML	0	3,0	1,0	3,2	3,1
CALB		0	2,3	0,8	2,7
FSC			0	2,4	3,3
BCL				0	3,1
DGL					0

Die strenge Konservierung des Ankers innerhalb der Familien, sowie die strukturell erhaltene Position des Ankers sprechen für einen modularen Aufbau der *oxyanion hole* bildenden Aminosäure und dem zugehörigen Anker. Um diese Hypothese zu bestätigen wurde das hydrophobe Modul der BCL gegen ein hydrophiles Modul wie in der Familie abH23.1 zu finden war, in zwei Schritten ausgetauscht. Zuerst wurde das *oxyanion hole* bildende L17 gegen ein Threonin ausgetauscht und dann der Anker L167 gegen die in der Familie abH23.1 als Anker vorkommenden Aminosäuren Aspartat, Glutamat, Asparagin und Glutamin ausgetauscht.

3.5.11 Expression der BCL und BCL Mutanten in E. coli

Der BCL Wildtyp, die Einfachmutante L17T und die Doppelmutanten L17T/L167D, L17T/L167E, L17T/L167N und L17T/L167Q wurde wie im Material und Methodenteil beschrieben hergestellt und in E. coli DH5α unter Kontrolle des hitzeinduzierbaren λ-Promotors exprimiert. Abbildung 34 stellt das coomassiegefärbte SDS Gel des überexprimierten BCL Wildtyps und der Einfachmutante L17T dar. Das Helferprotein wurde separat mittels des Plasmids pT-ompA-Δ70HpHis exprimiert. Die Doppelmutante L17T/L167E konnte nicht in einer aktiven Form gewonnen werden.

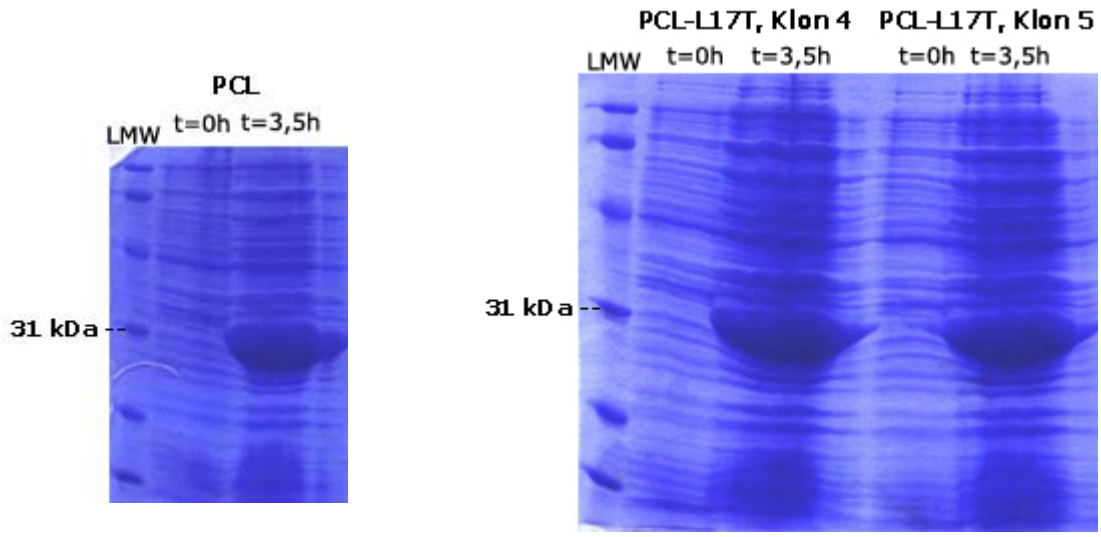


Abbildung 34: SDS-PAGE zum Nachweis der Expression von WT BCL (linke Abbildung) und BCL L17T.

3.5.12 Aktivitätsmessungen

Die spezifische Aktivität des BCL Wildtyps und der Mutanten wurde mit p-Nitrophenylpalmitat und Olivenöl gemessen wie im Methoden und Materialteil beschrieben. Die bestimmten Aktivitäten sind in Tabelle 8 aufgeführt. Der Austausch der *oxyanion hole* formenden Aminosäure L17 gegen ein Threonin resultierte in einer 200-fachen Verringerung der Aktivität gegen p-Nitrophenylpalmitat und einer 93-fachen Verringerung der Aktivität gegen Olivenöl, was der Inaktivierung der BCL entspricht. Durch den Austausch der Ankeraminosäure L167 gegen Asparagin, Aspartat und Glutamin wurde in allen Fällen eine Reaktivierung der Lipase erreicht. Die Doppelmutante L17T/L167N wies hierbei die höchste Reaktivierung auf. Gegenüber p-Nitrophenylpalmitat wurde ein 15-facher Anstieg, gegenüber Olivenöl ein 9-facher Anstieg der Aktivität beobachtet. Für die Doppelmutanten L17T/L167D und L17T/L167Q wurde ein 9-facher bzw. 6-facher Anstieg der Aktivität gegenüber p-Nitrophenylpalmitat, gegenüber Olivenöl ein 5-facher bzw. 8-facher Anstieg beobachtet.

Tabelle 8: Aktivitätsmessungen für BCL-WT und BCL Varianten gegenüber pNPP und Olivenöl

Mutante	BCA Abs562	Protein cont mg/ml]	pNPP dA/dt	Act. pNPP [U/100µL]	Spez. Act pNPP [U/mg]	Act. Olivenöl [U/mL]	Spec. Act. Olivenöl [U/mg]
WT	0,117	0,027	1,8	26,1	9699	0,3	11,15
L17T	0,595	0,147	0,05	0,73	49	0,02	0,12
L17T L167N	0,322	0,078	0,4	5,8	741	0,09	1,12
L17T L167D	0,405	0,099	0,3	4,35	439	0,06	0,58
L17T L167Q	0,225	0,053	0,11	1,63	302	0,05	0,91

3.5.13 Analyse der L17T/L167N Mutante

Für die Doppelmutante mit der höchsten Aktivität wurde ein Strukturmodell erzeugt. In Abbildung 35 ist der Vergleich des modulierten hydrophoben Moduls T17/N167 in BCL mit dem hydrophilen Modul in RML dargestellt. Die Methylgruppe des T17 fügt sich gut in die hydrophobe Umgebung der BCL. Diese kompensiert die hydrophoben Interaktionen des L17 im Wildtyp mit den Aminosäuren F52 und A163. Die Hydroxylgruppe des T17 ist in Richtung des Ankers N167 orientiert und kann mit dem Stickstoffatom (3,5 Å) oder dem Sauerstoffatom (3,7 Å) interagieren.

Im Vergleich zur Proteinstruktur der RML zeigte sich, dass sowohl die Orientierung der Seitenketten des Ankers als auch die der *oxyanion hole* bildenden Aminosäure übereinstimmen. Die Entfernung liegt zwar mit 3,5 Å über der einer idealen H-Brücke, ist jedoch im Bereich einer positiven Wechselwirkung.

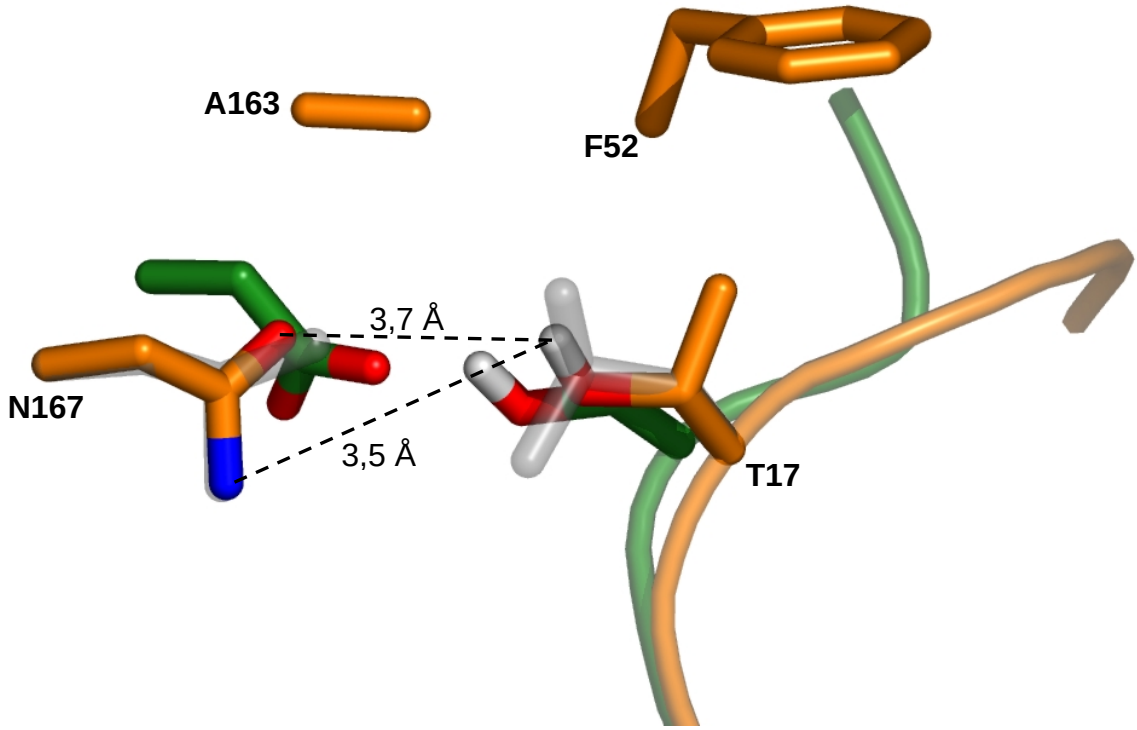


Abbildung 35: Vergleich der *oxyanion hole* Aminosäure, Anker Paare im Strukturmodell der BCL Doppelmutante L17T/L167N (orange) mit dem BCL Wildtyp (hellgrau) und der RML (grün). Die Methylgruppe des T17 der BCL Doppelmutante kann wie das L17 im WT der BCL mit den hydrophoben Aminosäuren F52 und A163 wechselwirken. T17 wird über den neuen Anker N167 stabilisiert, der in der Lage dem des Ankers der RML entspricht.

3.5.14 Familienspezifisches Primerdesign

Für die Superfamilie abH1 der GGGX Klasse wurden familienspezifische PCR Primer erstellt, die die Identifikation von neuen Mitglieder dieser Familie aus genomischer DNA oder Bodenproben ermöglichen. Neben familienspezifischen Motiven, die ausreichend konserviert sind um sich für ein solches PCR Primer Design zu eignen, ist ein zweites Kriterium für die Charakterisierung des isolierten Fragments wichtig. Das Fragment sollte eine minimale Länge von 20 Aminosäuren nicht unterschreiten, um mögliche Sequenzähnlichkeiten zu bekannten Proteinen signifikant bestimmen zu können. Unter Verwendung der in Block Maker implementierten Methode zur Bestimmung von konservierten Sequenzmotiven konnten neben dem für Serinhydrolasen bekannten GXSXG Muster und dem für diese Klasse typischen GGGX *oxyanion hole* Muster zwei weitere konservierte Bereiche identifiziert werden. Während für das GXSXG und das GGGX Motiv keine CODEHOP Hybrid Primer abgeleitet werden konnten, gelang dies für die anderen beiden Bereiche.

Diese beiden Bereiche umfassen den β-Strand β2 einschließlich den 4 N-terminal liegenden Aminosäuren (Blöcke $2M_3$, $2G_2$, siehe Anhang), sowie den Bereich von αA bis αC (Blöcke rM_5, rG_5+rG_6, siehe Anhang). Die für diese Bereiche durch CODEHOP abgeleiteten *Core* Regionen basieren auf zwei gut konservierten Aminosäuremustern. Für den β-Strand β2 waren dies die 4 N-terminal liegenden Aminosäuren, die das Motiv EDCL beschreiben. Für den Bereich von αA bis αC das Motiv FGGX. Das FGGX Motiv, das sieben Aminosäuren vor dem GXSXG Muster lokalisiert ist, liegt in etwa 100 Aminosäuren N-terminal zum EDCL Motiv.

Um die hohe Konservierung dieser beiden Motive zu verstehen wurden bekannte Proteinstrukturen dieser Superfamilie untersucht. Das FGGX Motiv bildet einen Loop, der die α-Helix αB mit dem zentralen β-Strand β5, an dessen C-terminalen Ende das katalytisch aktive Serin liegt, über einen engen Bogen miteinander verbindet. Dieses strukturelle Element erfordert die Flexibilität, die das GG Dipeptid bietet, womit sich dessen hohen Konservierung erklärt. Das Phenylalanin ist Teil eines hydrophoben Clusters, wogegen die variable Position X auf der Proteinoberfläche lokalisiert ist. Das EDCL Motiv geht dem β-Strand β2 voraus und ist in der Ausbildung einer Cysteinbrücke mit einem C-terminal lokalisierten Cystein involviert. Der durch die Cysteine umschlossene ω-Loop hat hierbei eine direkte Auswirkung auf die Substratspezifität dieser Enzyme. Abhängig von der Orientierung dieses Strukturelements und der Länge des ω-Loops wird der Zugang zur Bindungstasche bestimmt und nimmt somit Einfluss auf die Art der bevorzugten Substrate.

Da die Primer nur aus einem Teil der in der Superfamilie abH1 vorhanden Sequenzen abgeleitet wurden, wurde die statistische Signifikanz der für das Primerdesign verwendeten Motive EDCL und FGGX innerhalb der Superfamilie untersucht. Diese Superfamilie umfasst 321 Proteine. Die Analyse der 215 nicht putativen und nicht fragmentarischen Einträge ergab, dass das EDCL Motiv in 92% dieser Proteine vorkommt wogegen das FGGX Motiv variabler ist. Neben FGGD, der mit 58% am häufigsten auftretenden Variante, gibt es noch drei weitere Varianten (FGGN 26%, FGGE 8%, FNGD 5%). Da die Superfamilie abH1 ein breites Spektrum an verschiedenen Organismen umfasst, war es notwendig, dass diese auch durch die Primer abgedeckt wurden. Für das EDCL Motiv ist dies gegeben, wogegen für das FGGX Motiv die 4 häufigsten Motive untersucht wurden. Die beiden Motive FGGD und FGGN decken gemeinsam 84% des Sequenzraums der abH1 Proteineinträge ab und repräsentieren die gesamte taxonomische Vielfalt dieser Familie. Die Motive FGGE und FNGD treten nur in Proteinen aus der Klasse Anthropoda auf.

Da Block Maker Sequenzen, die eine sehr geringe Ähnlichkeit aufweisen von der Bestimmung der Sequenzblöcke ausschließt, muss gewährleistet sein, dass in den Sequenzblöcken, die zum Design der reversen Primer genutzt wurden, die gefunden Varianten für das FGGX Motiv repräsentiert sind. Sowohl in rM_5 als auch in rG_5 sind die Varianten FGGD und FGGN enthalten. Diese wurden von CODEHOP zur Konstruktion der Primer beide berücksichtigt. Die Position X des FGGX Motivs ist im *Core* Bereich der Primer rBS und rHS durch das Codon RAC repräsentiert. Dies kodiert sowohl für Aspartat als auch Asparagin.

3.5.15 Experimentelle Validierung der familienspezifischen Primer

Die Eignung der *in silico* erzeugten familienspezifischen Primer für die praktische Isolierung neuer Carboxylesterasen, wurde durch Jutta Schmitt im Experiment überprüft. Im folgenden sind die von Jutta Schmitt erhaltenen Ergebnisse zusammengefasst.

Für die Etablierung des Screening Systems wurde *Bacillus subtilis* als Modellorganismus gewählt, da dessen pNBA einen Vertreter der Superfamilie abH1 darstellt und das komplette Genom sequenziert wurde. Neben den degenerierten familienspezifischen Primern wurden zur Kontrolle Primer für die spezifische Amplifikation des pNBA Fragments in einer Touchdown PCR mit Annealingtemperaturen von 60 bis 45°C verwendet.

Eine einzige scharfe Bande der erwarteten Länge von 320 bp wurde für alle Primerpaare beobachtet, mit Ausnahme des degenerierten Primers v2 in Kombination mit den Primern rBS und rHS

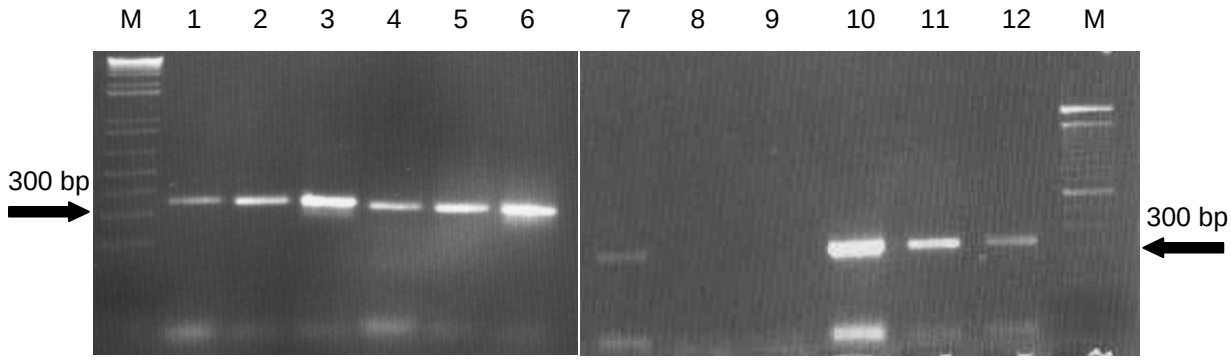


Abbildung 36: Amplifikation aus der genomischen DNA für *Bacillus subtilis* mit verschiedenen Primerkombinationen: (1) c-EDCLfw/ rBS, (2) c-EDCLfw/rHS, (3) c-EDCLfw/r-pNBA, (4) v-pNBA/rBS, (5) v-pNBA/rHS, (6) v-pNBA/r-pNBA, (7) v2/rpNBA, (8) v2/rHS, (9) v2/rBS, (10) v1/r-pNBA, (11) v1/rHS, (12) v1/rBS

Es wurde kein PCR Produkt mit unerwarteter Länge erhalten. Die 320 bp langen PCR Produkte wurde mittels des TOPO TA Kit kloniert. Für jede Klonierung wurden mindestens 3 Klone sequenziert. Für alle Sequenzanalysen wurden die den degenerierten Primern entsprechenden Sequenzbereiche entfernt. Homologiesuchen gegen Genbank mit BLAST ergaben, dass alle PCR Produkte sehr hohe Ähnlichkeiten mit pNBA des *Bacillus subtilis* Stamms B8079[108] aufwiesen.

Um die Anwendbarkeit der Primer für das in vitro Screening verschiedener Organismen weiter zu untersuchen wurden PCR Reaktionen mit der genomischen DNA der Organismen *Bacillus megaterium* und *Bacillus firmus* durchgeführt. Für beide Organismen wurde bisher keine Carboxylesterasen beschrieben.

Für *Bacillus firmus* wurde eine scharfe Bande mit 345 bp Länge mit den Primerpaaren c-EDCLfw/r-pNBA, c-EDCLfw/rHS und v1/rHS erhalten, kloniert und sequenziert. Die Sequenzierung von 18 Klonen ergab eine sehr hohe Ähnlichkeit mit der Carboxylesterase B aus *Xanthomonas campestris* und *Xanthomonas acitri*, der Carboxylesterase aus *Bacillus sp.* BP-23 und pNBA aus *Bacillus subtilis*.

Für *Bacillus megaterium* ergaben sich PCR Produkte unterschiedlicher Länge mit 103 bp, 170 bp, 428 bp und 467 bp. Klonierung und Sequenzierung dieser PCR Produkte ergab, dass keine dieser Sequenzen Ähnlichkeit mit Carboxylesterasen aufwies, jedoch verwandt waren mit anderen Proteinen wie dem PBD Transportsystem Protein aus *Bacillus anthracis*, einer hypothetischen Aminotransferase aus *Bacillus firmus*, dem yhdE Gen aus *Bacillus subtilis* und *Bacillus halodurans* und der Malatdehydrogenase aus *Bacillus subtilis*.

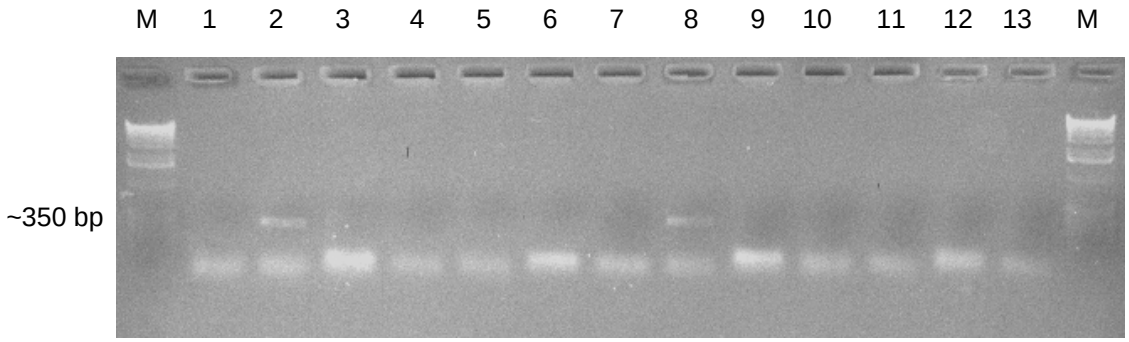


Abbildung 37: Amplifikation aus genomischer DNA für *Bacillus firmus*. Für die Primerkombinationen c-EDCLfw/rHS (2), c-EDCLfw/r-pNBA (3), v1/rHS (8) wurde ein PCR Produkt bei 345 bp Länge erhalten.

Es ist bekannt, dass einige Organismen und im speziellen Pilze, multiple Lipase/Esterase Gene mit unterschiedlichen Aktivitäten enthalten können. Deshalb wurde versucht die Primer zur Bestimmung multipler Carboxylesterase Gene in Organismen zu verwenden. Für die methylotrophe Hefe *Candida rugosa* sind mindestens 7 Lipasegene bekannt.[109] Die Familie Candida rugosa ist zwar kein Mitglied der Superfamilie GGGX Carboxylestaerses, sie ist aber dennoch nahe mit diesen verwandt und enthält ebenso das EDCL sowie FGGX Motiv. PCR Reaktionen mit allen Primerpaaren wurde mit genomische DNA der *Candida rugosa* durchgeführt, wobei die Primer v1 und v2 in Kombination mit rBS und rHS zu einem 350 bp lange PCR Produkt führten. Alle anderen Primerkombinationen führten zu keinem PCR Produkt. Die Klonierung und Sequenzierung von 62 PCR Produkten mit anschließenden Sequenzvergleichen mittels BLAST Homologiesuchen ergab, dass nicht alle fünf bekannten Isoformen gefunden wurden. 21 der 62 Klone waren hoch ähnlich zum Gen Lip1, und jeweils ein Klon entsprach den Genen Lip4 und Lip5. Die übrigen 39 Klone zeigten geringere aber signifikante Ähnlichkeiten zu den bekannten *Candida rugosa* Lipasegenen.

Ein weiterer Pilz, *Aspergillus nidulans*, wurde auf das Vorhandensein von Carboxylesterasegenen gescreent. PCR Reaktionen mit genomischer DNA, die aus Sporen isoliert wurde, ergaben für die Primer fEDCL1 und fEDCL2 in Kombination mit rFGGX1 und rFGGX2 Fragmente unterschiedlicher Länge. Sequenzvergleiche von 16 sequenzierten Klonen zeigte, dass die verschiedenen Primerpaare unterschiedliche PCR Produkte ergaben. Das Primerpaar fEDCL2/rFGGX2 ergab 4 Klone mit hoher Ähnlichkeit zur *Bacillus subtilis* Carboxylesterase B (424 bp Fragment) wogegen 4 weitere Klone keine Ähnlichkeit zu Carboxylesterasen aufwiesen. Für das Primerpaar fEDCL1/rFGGX2 wurden 6 Klone mit einem 381 bp langen Fragment mit sehr hoher Ähnlichkeit zu Lipasen aus *Candida rugosa* und *Geotrichum candidum* bestimmt. Zwei Klone mit einem 366 bp langen Fragment hatten hohe Ähnlichkeit mit einem murine 62,1 kDa hypothetischen Protein und dem humanen

Neuroligin Y. Diese Ergebnisse deuten stark darauf hin, dass in *Aspergillus nidulans* Carboxylesterasen vorhanden sind.

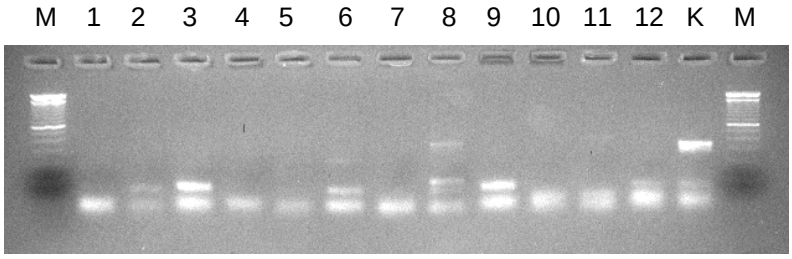


Abbildung 38: Amplifikation aus genomischer DNA für *Bacillus nidulans*. Für die Primerkombinationen: (1) c-EDCLfw/ rBS, (2) c-EDCLfw/rHS, (3) c-EDCLfw/r-pNBA, (4) v-pNBA/rBS, (5) v-pNBA/rHS, (6) v-pNBA/r-pNBA, (7) v2/rpNBA, (8) v2/rHS, (9) v2/rBS, (10) v1/r-pNBA, (11) v1/rHS, (12) v1/rBS, (K) *Bacillus subtilis* Fragment

4 Diskussion

Zielsetzung dieser Arbeit war zu zeigen, dass die systematische Analyse einer Proteinfamilie zum Verständnis der Sequenz-Struktur-Funktionsbeziehung entscheidend beiträgt. Mit Hilfe des für diese Anwendung speziell entwickelten *Data Warehouse* Systems, konnte dies am Beispiel der α/β-Hydrolasen erfolgreich demonstriert werden. Die Handhabung der großen Menge an Sequenz- und Strukturdaten, die für diesen Ansatz benötigt werden, wurde durch das *Data Warehouse* effizient gestaltet und führte zum Aufbau der *Lipase Engineering Database* (LED). Die LED erwies sich als geeignete Grundlage zur systematischen Untersuchung der α/β-Hydrolasen. Die Analyse des Sequenzraums der α/β-Hydrolasen führte zur Klassifikation wohl definierte Familien. Die Qualität dieser Klassifikation war entscheidend für die Entwicklung sequenzbasierte Deskriptoren wie Profil HMMs oder familienspezifischer Primer. Während die Analyse der Sequenzen zur Beschreibung von Eigenschaften der gesamten Proteinfamilie diente, trug die Untersuchung einzelner Strukturen zum Verständnis der molekularen Funktionsweise bei. Die detaillierte Beschreibung der Fettsäurebindungstaschen verschiedener Esterasen und Lipasen konnte die individuellen Substratspezifitäten auf molekularer Ebene erklären. Aber nur die Kombination der Sequenzeigenschaften mit der strukturellen Beschreibung der molekularen Funktionsweise führte zu einem tiefern Verständnis der Sequenz-Struktur-Funktionsbeziehung. So konnten die bisher unbekannten aber funktionell relevanten Ankeraminosäuren, die für die Stabilisierung der *oxyanion hole* Aminosäuren entscheidend sind, identifiziert werden und die für die Faltung des α/β-Hydrolase Folds wichtigen Elemente, der sogenannte Faltungsnukleus, bestimmt werden.

4.1 Sequenzanalyse

Die vergleichende Sequenzanalyse hat sich als Methode zur deduktiven Beschreibung familienspezifischer Eigenschaften der α/β-Hydrolasen bewehrt. Die große Menge an bekannten Sequenzdaten stellte genügend Informationsgehalt für eine signifikante statistische Analyse bereit. Grundlage dieser Analyse waren wohl definierter Sequenzfamilien, die über die automatisierte Klassifizierung mit anschließender manueller Überarbeitung durch den Datenbank *curator* effektiv bestimmt wurden. Innerhalb dieser Familien war es möglich zuverlässig funktionell relevante Annotationen auf nicht annotierte Proteinsequenzen zu übertragen, Signaturen für einzelne Familien abzuleiten, die für die Identifikation neuer

α/β-Hydrolasen aus genomischer DNA geeignet sind und über die Erstellung von HMMs als familienspezifische Deskriptoren die phylogenetische Verwandtschaftsbeziehungen innerhalb der α/β-Hydrolasen nachzuvollziehen.

4.1.1 Annotation und Klassifikation

Die Information über Proteinsequenzen sowie die Anzahl und Größe von Sequenzdatenbanken steigt rasch an. Neben den reinen Sequenzdaten enthalten Datenbanken wie Swiss-Prot Informationen über Ähnlichkeiten zu anderen Proteinen, Annotationen funktionell relevanter Aminosäuren und bekannte Mutationen. Jedoch muss für jeden neuen Eintrag diese Information manuell durch den *curator* der Datenbank editiert werden. Dies stellt den zeitaufwendigsten Teil der Verwaltung von Datenbanken dar. So enthält die automatisch translatierte EMBL Datenbank (TrEMBL) 1377572 Einträge (August 2004) die vollständig annotierte Swiss-Prot jedoch nur 158010 Einträge.[110]

Der große Vorteil von Datenbanken wie Swiss-Prot ist das breite Spektrum an bereitgestellten biologischen Informationen und ihre standardisierte Nomenklatur. Auf der anderen Seite führte die über das World Wide Web leichte Verfügbarkeit zu einer Vielzahl von biologischen Datenbanken, die weitere Informationen bereitstellen wie Klassifikation von Enzymen auf Sequenzebene (SYSTERS,[91] Pfam,[25] ProDom[111]) oder über die Proteinstruktur (SCOP,[63] CATH[7]) sowie Motiv Datenbanken, die enzymfamilienspezifische Muster bereitstellen (Prosite,[18] PRINTS[22]). Suchmaschinen wie SRS[51,112] oder Entrez[53] haben sich als hilfreiche Werkzeuge zur Integration dieser heterogenen Daten erwiesen.

Während das Ziel dieser Datenbanken darin liegt den gesamten Sequenzraum abzudecken, konzentriert sich die LED auf eine Proteinfamilie, die der α/β-Hydrolasen, mit dem Ziel alle notwendigen Informationen für die Familie zu integrieren, die für das Verständnis der Sequenz-Struktur-Funktionsbeziehung notwendig sind. Eine verwandte Datenbank stellt ESTHER[31,32] dar, die mit Focus auf Acetylcholinesterasen sich auch der α/β-Hydrolase Fold Superfamilie widmet.

Die LED stellt auf globale Multisequenz Alignments aufbauend eine Klassifikation aller α/β-Hydrolasesequenzen in homologe Familien und Superfamilien bereit. Die Klassifikation der LED wurde mit der Klassifikation der SYSTERS Datenbank[91] verglichen, die auf iterierten BLAST Suchen basiert und alle Proteineinträge der *Swiss-Prot* und TrEMBL Datenbanken enthält. SYSTERS klassifiziert ähnlich wie die LED in Superfamilien und homologe Familien ein. SYSTERS homologe Familien, auch Cluster genannt, werden über eine *single-linkage* Cluster Methode bestimmt und sind aufgrund des Grads der Konsistenz in

drei Typen aufgeteilt, perfekte oder *single* Sequenz Cluster (P, S), *nested* Cluster (N) und *overlapping* Cluster (O). Im Allgemeinen entsprechen die wohl definierten P und S Cluster homologen Familien der LED. Im Falle der O und N Cluster ist jedoch häufig eine Aufteilung in mehrere homologe Familien zu beobachten wie für die Superfamilien abH1, abH4, abH14, abH15, abH20, abH23, abH31, abH32, abH33 abH34 und abH36 deren homologe Familien sich aus jeweils einem N oder O SYSTERS Cluster ableiten. Diese Unterschiede lassen sich durch die unterschiedlichen Methoden der Klassifikation erklären. Während SYSTERS über lokale Ähnlichkeiten eine Clusterzugehörigkeit definiert, werden in der LED homologe Familien über globale Multisequenz Alignments klassifiziert. Vergleicht man jedoch die Ebene der Superfamilien zeigt sich, dass SYSTER hier eine feinere Einteilung aufweist. So sind in den Superfamilien abH1, abH7, abH8, abH9, abH11 und abH15 jeweils mehrere SYSTERS Familien in einer Superfamilie vereint. Diese Unterschiede basieren auf dem relativ geringen Schwellenwert, den SYSTERS für die Klassifikation der Superfamilien heranzieht, während in der LED die Zugehörigkeit zu einer Superfamilie über die Konservierung der katalytischen Triade im Multisequenz Alignment abgeleitet wird.

Die Proteinfamilien Datenbank Pfam,[25] klassifiziert Proteine über Hidden Markov Modelle für manuell überarbeitete Alignments. Diese Alignments repräsentieren Proteindomänen, so dass ein Protein Mitglied verschiedener Familien sein kann. In Pfam sind zwei Klassen von Familien definiert: (1) PfamA Familien, die die HMMs der Pfam repräsentieren und (2) PfamB Familien, einem automatisch erzeugten Teilbereich der Pfam, die sich aus der ProDom Datenbank[111] ableitet. 30 Superfamilien der LED lassen sich 16 PfamA Familien zuordnen. Die 7 verbleibenden Superfamilien zeigen signifikante Ähnlichkeit zu verschiedenen PfamB Familien. Somit weist die LED eine feinere Einteilung als Pfam auf. Die PfamA Familie Abhydrolase_1 umfasst z.B. die Superfamilien abH7-abH12 und abH14. Die Pfam Datenbank bietet keine phylogenetischen Distanzen zwischen den PfamA Familien an, aus denen der Verwandtschaftsgrad zwischen den Familien abgeleitet werden könnte. Überträgt man die PfamA Familien auf die Topologie des distanzbasierten Baums der α/β-Hydrolasen zeigt sich jedoch, dass PfamA Familien nur Superfamilien der LED in sich vereinigen, die phylogenetisch verwandt sind.

Für bakterielle Lipasen haben Arpigny und Jäger[113] eine Klassifikation in 8 Klassen (I-VIII) anhand von Sequenzmotiven beschrieben. Während die größte Klasse I Vertreter der phylogenetisch nicht direkt verwandten Superfamiliengruppen abH15, abH16, abH17 sowie abH24 vereint, stellen die Klassen III-VII einzelne oder nah verwandte Superfamilien dar. Die

Klassen II und VIII sind nicht in der LED enthalten, da diese nicht dem α/β-Hydrolase Fold sondern dem Flavodoxin-like bzw. β-Lactamase Fold angehören.

Der Vergleich und die Analyse von Proteinfamilien erlaubt eine konsistente und schnelle Annotation konservierter Aminosäuren, wie anhand der α/β-Hydrolasen gezeigt werden konnte. Für alle Superfamilien, für die Proteinstrukturen verfügbar waren, konnte die katalytische Triade sowie das *oxyanion hole* identifiziert werden. Über Sequenzvergleiche konnten so diese funktionellen Aminosäuren für 85% der homologen Familien und 93% aller LED Proteineinträge zugewiesen werden. Sequenzalignments und Strukturüberlagerungen der homologen Familie abH15.2 zeigte, dass das katalytisch aktive Aspartat in 4 von 5 Swiss-Prot Einträgen falsch annotiert war. Für die homologen Familien abH11.2, abH7.2, abH4.1 und abH4.2 war in Swiss-Prot ein falsches Histidin als katalytisch aktiv verzeichnet. Dieses annotierte Histidin liegt im N-terminalen Bereich der Sequenz und nicht wie erwartet C-terminal zum katalytisch aktiven Nucleophil. Für die Superfamilie abH4 konnte aufgrund vorhandener Proteinstrukturen das katalytisch aktive Histidin bestimmt werden und es wurde gezeigt, dass das katalytisch aktive Histidin mit dem sehr gut erhaltenen Histidin des *oxyanion holes* vertauscht wurde. Für die Superfamilien abH7 und abH11 sind zwar keine Proteinstrukturen bekannt, jedoch ergab die HMM Analyse, dass diese Familien verwandt sind mit den Superfamilien abH8 und abH9, die beide bekannte Proteinstrukturen enthalten. Aufgrund lokaler Ähnlichkeiten konnte so auch für diese Familien festgestellt werden, dass das als katalytisch aktiv beschriebene Histidin mit dem Histidin des *oxyanion holes* vertauscht wurde.

Multisequenz Alignments für Superfamilien und homologe Familien ermöglichen Informationen eines Familienmitglieds auf alle anderen Mitglieder der jeweiligen Familie zu übertragen. So kann Sekundärstrukturinformation einer α/β-Hydrolase mit bekannter Proteinstruktur auf Sequenzen mit unbekannter Struktur abgebildet werden, funktionell relevante Aminosäuren wie die katalytische Triade oder das *oxyanion hole* können identifiziert werden und die Effekte von Mutationen können mit der Proteinsequenz und Struktur korreliert werden. Für die Lipasen aus *Rhizopus delemar* und *Rhizomucor miehei* wurden mehrere Mutationen und ihre experimentell beobachteten Effekte beschrieben. Durch die Analyse der Konservierungsmuster innerhalb der homologen Familien konnten diese experimentellen Beobachtungen verstanden werden, wie die Auswirkung auf die katalytische Aktivität[114] oder die Kettenlängenspezifität, wie hier gezeigt wurde.

Da das zentrale α/β/α Motiv das einzige strukturelle Element darstellt, das alle α/β-Hydrolasen gemeinsam haben, ist eine Überlagerung aller α/β-Hydrolasen nicht

eindeutig möglich. Die hier beschriebene Strategie zur Überlagerung ist für die Substratbindungstasche und die katalytische Maschinerie optimiert. So kann die Position und Orientierung des tetrahedralen Kohlenstoffatoms des Substratübergangszustandes und dessen Interaktion mit dem Enzym verglichen werden. Für experimentell bestimmte Protein-Inhibitor Komplexe der RML und CRL (Phosphat bzw. Phosphonat als Inhibitor) wurde die Überlagerung der LED Struktureinträge verglichen mit der der FSSP Datenbank,[115] die globale Struktur Alignments bereitstellt. Die Abweichung des Phosphoratoms und seiner vier Liganden betrug 1 Å für die LED Struktureinträge während diese für die FSSP Überlagerung 1,5 Å beträgt. Somit ist die Überlagerung der LED Struktureinträge für das Aktive Zentrum optimiert, obwohl sich die Geometrie und physikalisch-chemischen Eigenschaften dieser beiden Lipasen deutlich unterscheiden.

4.1.2 Familienspezifische Primer

Das Auffinden von neuen Proteinen über PCR basierte Suchmethoden hat sich bereits als Alternative zur aktivitätsbasierten Suche in Bibliothek bewehrt.[116] Um diese Methode auf Proteinfamilien anwenden zu können, müssen Sequenzmotive bekannt sein, die geeignet sind um proteinfamilienspezifische CODEHOP Hybridprimer zu generieren. Die hier beschriebenen Untersuchungen deuten darauf hin, dass durch die Variation des Grads der Degenerierung dieser Primer im Vergleich zu spezifischen Suchmethoden eine Optimierung der Abdeckung für die Identifikation von neuen Proteinen erzielt werden kann.

Für nahe verwandte Proteine können Primer mit einer geringen Degenerierung konstruiert werden, die ein sehr spezifisches Auffinden von homologen Proteinen erlauben. Jedoch lässt sich dieser Ansatz nur innerhalb eines schmalen Spektrums an Organismen anwenden, wie z.B. für die Entdeckung von Chitinasen aus Bakterien und Streptomyceten.[117]

Bell et al. verwendeten die Methode der degenerierten Hybridprimer für die Detektion von neuen Lipasen.[118] Über den Vergleich der Motive um das katalytisch aktive Serin und die Region des *oxyanion holes* identifizierten sie vier Gruppen von Lipasen, wobei sie das Primerdesign auf die größte Gruppe konzentrierten. Diese Gruppe umfasste nur Lipasen der GX Klasse aus fünf Familien der LED (abH8.9, abH14, abH15.2, abH18.1, abH23.2). Da diese Gruppe fünf Familien der hoch diversen GX Klasse umfassen, wiesen die daraus abgeleiteten Primer einen hohen Grad der Degenerierung auf. Die Amplifikation von DNA aus Biomasse mit diesen Primern führte, neben einer Vielzahl nicht weiter untersuchter Fragmente, zur Identifikation einer neuen Lipase. Die drei weiteren von Bell et al. definierten Gruppen umfassten entweder ein enges phylogenetisches Spektrum an Lipasen (zwei GX Superfamilien) oder aber eine Mischung verschiedener Lipasen sowohl aus der GX als auch

GGGX Klasse. Das lässt darauf schließen, dass Primer die aus diesen Gruppen abgeleitet werden, entweder zu einem engen Screening mit einer geringen Abdeckung wie im Fall von Williamson et al. führt,[117] oder zu einer unspezifischen Amplifikation von verschiedenen Sequenzfragmenten. Im Vergleich zur Klassifikation nach Bell et al., ist die auf Sequenzähnlichkeit basierende Klassifikation der LED sehr gut geeignet, um Motive zu identifizieren, für die enzymfamilienspezifische Primer abgeleitet werden können. Dies hilft die Abdeckung des Screenings im Vergleich zur Spezifität zu optimieren. Nicht nur der Grad der Konservierung ist für die Wahl der Motive entscheidend, sondern auch Informationen über die strukturellen und funktionellen Eigenschaften des Motivs müssen berücksichtigt werden.

Obwohl die bekannten Motive des katalytisch aktiven Serins (GXSXG) und des *oxyanion holes* funktionell relevant sind, ist das häufige Vorkommen dieser Motive und der hohe Grad der Degenerierung aufgrund des hohen Glycingehalts für ein selektives Screening ungeeignet. Im Gegensatz dazu erlauben die aus der LED abgeleiteten Motive ein spezifischeres Screening. Die Analyse von Multisequenz Alignments zeigte eine höher Konservierung dieser Muster im Vergleich zu denen des katalytisch aktiven Serins und des *oxyanion holes*. Proteinstrukturanalysen ergaben, dass die hohe Konservierung dieser Motive auf deren strukturelle Funktion für die GGGX Klasse zurückzuführen ist. Das FGGD Motiv bildet einen engen Bogen zwischen der α-Helix αB und dem β-Strand β5 die innerhalb der α/β-Hydrolasen streng konserviert sind. Die Funktion des EDCL Motivs ist, wie von Henke et al. gezeigt werden konnte, sogar an die Substratspezifität dieser Enzymklasse gekoppelt.[119] Das Cystein des EDCL Motivs bildet mit einem C-terminal liegenden Cystein eine Disulfidbrücke, die den sogenannten ω-Loop einschließt. Die Orientierung dieses strukturellen Elements reguliert den Zugang zur Bindungstasche und spiegelt so den Charakter der bevorzugten Substrate wieder. Henke et al. beschrieben drei verschiedene Typen des ω-Loops, abhängig von dessen Länge, wobei diese mit der Substratspezifität der Enzyme korrelieren: 1. lange, flexible Loops bilden das Lid der Lipasen wie im Fall der Lipase aus *Candida rugosa*, 2. lange, fixierte Loops wie im Fall der Cholinesterasen und 3. Esterasen, die sich in mittlere und kurze Loops aufteilen. Die Screeningergebnisse für die EDCL Primer zeigten, dass positive Treffer nur für den ω-Loop Typ 1, wie im Fall der *Candida rugosa*, und den kurzen ω-Loops des Typs 3 der Bacillus Stämme gefunden wurden. Jedoch konnten keine positiven Treffer für Typ 2 (humane Acetylcholinesterase) und Typ 3 Enzyme mit mittlerem ω-Loop (Schweineleber Lipase) identifiziert werden. Dieses Ergebnis korreliert auch mit der Anzahl der nicht übereinstimmenden Positionen zwischen dem

Consensus *Clamp* der EDCL Primer und den *Template*-Sequenzen. Diese Region ist für *Candida rugosa* und die Bacillusstämme ähnlicher als für die nicht amplifizierten PLE und huAChE. Dies lässt vermuten, dass der Consensus *Clamp* des EDCL Motivs für ein substratspezifisches Screening von Carboxylesterasen genutzt werden kann.

Die Sequenzanalyse der homologen Familien der Superfamilie abH1 zeigte, dass neben dem ω-Loop auch der Bereich, der durch die Motive EDCL und FGGX eingeschlossen wird, für eine Zuordnung von Sequenzen genutzt werden kann. Da diese homologen Familien unterschiedliche Substratspezifitäten wiederspiegeln, lassen sich somit unbekannte Enzyme klassifizieren und Substratspezifitäten vorhersagen. Die aus *Aspergillus nidulans* amplifizierten Fragmente zeigten Ähnlichkeit zur Lipase aus *Candida rugosa*, der Carboxylesterase B aus *Bacillus subtilis* bzw. dem humanen Neuroligin Y.

Obwohl die Motive, die als *Core* Region gewählt wurden, das gesamte taxonomische Spektrum an Organismen der Familie abH1 abdecken, konnte die abgeleiteten Primer nicht erfolgreich auf alle Organismen angewandt werden. Während für Bakterien und Pilze das Screening erfolgreich war, konnten für höhere Organismen und Streptomyceten bekannte Carboxylesterasen nicht amplifiziert werden. Die Sequenzanalyse der putativen Carboxylesterasen CAB55678 und CAC37455 aus *Streptomyces coelicolor* ergab, dass das EDCL Motive der *Core* Region für diese Proteine gegen EDIL bzw. EDYL ausgetauscht waren. Dies führte zu einer bzw. zwei nicht übereinstimmenden Nucleotiden in der *Core* Region und kann in Verbindung mit den nicht idealen *Clamp* Regionen zum fehlenden PCR Produkt geführt haben.

Auch konnten die bekannten Gene für Leberesterase und Acetylcholinesterase aus den höheren Organismen *Sus scrofa* und *Homo sapiens* nicht amplifiziert werden, obwohl diese die Motive EDCL und FGGD enthalten. Die Untersuchung der Consensus *Clamp* Regionen N-terminal zu EDCL zeigte, dass eine erfolgreiche Amplifikation mit einer übereinstimmenden *Clamp* Region korrelierte. Dies führt dazu, dass die Hypothese die *Clamp* Region hätte einen geringen Einfluss auf die Amplifikation, überdacht werden sollte. Da in diesem Fall die *Clamp* Region die Substratspezifität der Carboxylesterasen wiederspiegelt, sollte die *Clamp* sogar zum Optimieren des Screening genutzt werden können. Wie hier beschrieben scheint es möglich zu sein ein substratspezifisches Screening für Carboxylesterasen der Familie abH1 etablieren zu können, das auf der *Clamp* Region, die den ω-Loop repräsentiert, basiert.

Morant et al.[11] beschrieben ebenso die Limitationen des Screenings mit Consensus Hybridprimer. Es war ihnen nicht möglich erwartete Fragmente für P450 Gene aus einer

Pflanzen cDNA Bibliothek zu erhalten. In diesen Fällen stimmten 2 von 16, 4 von 16 bzw. 4 von 11 Positionen nicht überein.

4.2 Strukturanalyse

Während die Sequenzanalyse zur Bestimmung familienspezifischer Eigenschaften eingesetzt wurde, konnte die Strukturanalyse einzelner Individuen zum Verständnis der molekularen Funktionsweise dieser α/β-Hydrolasen beitragen. So konnte durch die Untersuchung der Fettsäurebindungstasche für acht α/β-Hydrolasen die unterschiedlichen Substratspezifitäten nachvollzogen werden. Mit diesen individuell abgeleiteten Modellen der Substratspezifität war es möglich auf molekularer Ebene den Effekt einzelner Mutation auf die Substratspezifität zu erklären. Dieser Ansatz ist auch für das rationale Proteindesign von großem Vorteil, da Mutationen zur gezielten Änderung der Substratspezifität vorhergesagt werden können.

4.2.1 Anatomie der Lipasen

In der α/β-Hydrolase Foldfamilie existieren zwei große Untergruppen, die auf die Hydrolyse von Esterbindungen spezialisiert sind. Esterasen, die lösliche Ester spalten und Lipasen, die unlösliche, langkettige Triglyceride hydrolysieren. Esterasen und Lipasen sind jedoch bemerkenswert ähnlich in ihrer Struktur und Funktion.[5,63] Sie nutzen einen gemeinsamen enzymatischen Mechanismus, wobei das katalytisch aktive Zentrum am Grund einer tiefen, elliptischen Bindungstasche liegt. Dort binden Estersubstrate mit der Fettsäure-Alkoholachse parallel zur Längsachse der Bindungstasche. Die Bindungstasche hat in allen hier untersuchten Serinesterasen und Lipasen mit 4-4,5 Å die selbe Breite und hat die geeignete Größe um Estersubstrate zu fixieren. Die Länge der Acylbindungstasche variiert von 3,5 Å in der Acetylcholinesterase und Bromoperoxidase bis 22 Å in der *Candida rugosa* und *Rhizomucor miehei* Lipase. Somit können die hier untersuchten Hydrolasen anhand der Größe der Acylbindungstasche in zwei Klassen eingeteilt werden: (1) Esterasen, die Substrate mit kurzen Acylresten wie Acetyl, Propionyl und Butyryl hydrolysieren und (2) Lipasen, die mittlere bis lange Fettsäureketten akzeptieren.

Substratspezifität wird vermittelt über die Form der Acylbindungstasche. Esterasen haben eine kleine Acylbindungstasche, die optimal für die Bindung der Acylreste der Substrate geeignet ist. Die Verkleinerung der Bindungstasche würde zu sterischen Wechselwirkungen mit dem Substrat führen, wogegen eine Vergrößerung der Bindungstasche zu einer suboptimalen Bindung des Substrats führen würde und so eine Verringerung von k_{cat}/K_m

verursachen würde, was für eine Mutante der Acetylcholinesterase gezeigt wurde.[120] Lipasen besitzen eine lange, hydrophobe Fettsäurebindungsstelle, die in einer Spalte am Rand der trichterförmigen Bindungstasche oder in einem Tunnel lokalisiert ist. Cutinase hat eine deutlich geringer hydrophobe Fettsäurebindungsstelle innerhalb der Bindungstasche und stellt somit einen Vertreter der Esterasen dar. Die Form der Bindungstasche ist jedoch typisch für Lipasen. Somit weist die Bindungstasche der Cutinase eine Kombination der Eigenschaften von Lipasen und Esterasen auf.

Dieses Ergebnis führt zu zwei Hypothesen, die im weiteren genauer beschrieben werden: (1) nur eine kleine, begrenzte Anzahl von Aminosäuren formen die Fettsäurebindungstasche. Der Austausch dieser Aminosäurepositionen sollte eine gezielte Änderung der Kettenlängenspezifität ermöglichen. (2) Länge und Hydrophobizität dieser Bindungstaschen sollten mit den Profilen der Fettsäurekettenlängen der entsprechenden Lipasen korrelieren. Somit sollte sich die Kettenlängenspezifität anhand der Struktur einer Lipase vorhersagen lassen, falls die Kettenlänge des untersuchten Substrats nicht länger als die Fettsäurebindungsstelle der Lipase ist. Denn sobald die Länge des Substrats die der Bindungstasche überschreitet, kann die Konformation der Fettsäurekette nicht mehr vorhergesagt werden, da diese auf deren hydrophoben Oberfläche des Proteins bindet, in die hydrophobe Grenzschicht des Substrats ragt, oder sogar zurück in die Bindungstasche binden könnte. Im Falle der Strukturaufklärung solcher Komplexe, kann die Konformation des externen Teils der Substratkette durch Kristallkontakte beeinflusst sein.

4.2.2 Mutationen der Fettsäurebindungsstelle der Lipase aus *Rhizomucor miehei*

In Tabelle 9 sind experimentelle Daten für Mutanten der Lipase aus *Rhizopus delemar* und *Rhizopus oryzae* mit veränderter Kettenlängenspezifität dargestellt. Im Vergleich mit der modellierten Substratbindung innerhalb der Fettsäurebindungsstelle (Abbildung 31) zeigt sich, dass alle experimentellen Beobachtung anhand der Position der ausgetauschten Aminosäure und deren Wechselwirkungen mit der Fettsäurekette erklärt werden können. V205 liegt am Grund der Bindungstasche nahe der C4 Position der Fettsäurekette. Der Austausch gegen ein hydrophiles Threonin führt zu einer 5-10fachen Verringerung der Aktivität. Der Austausch von F111, das nahe der C6 Position liegt, gegen ein Tryptophan erhöht die relative Spezifität gegenüber kurzkettigen Triglyceriden. L208 und F94 liegen am Rand des hydrophoben Spalts nahe den Positionen C10 und C12. Der Austausch gegen sterisch anspruchsvollere und hydrophilere Aminosäuren verringert die Aktivität und erhöht die relative Spezifität gegenüber Fettensäuren mit kurzen und mittleren Kettenlängen. F213 liegt am Ende des hydrophoben Spalts nahe der C16 Position. Der Austausch dieser

Aminosäure gegen ein sterisch anspruchsvolleres und hydrophiles Tyrosin verringert die relative Spezifität gegenüber Öl- und Stearinsäure.

Als Strategie für die Vorhersage von Mutanten mit einer Kettenlängenspezifität unterhalb einer gegebenen Länge lässt sich somit feststellen, dass die Bindung längerer Ketten durch die Erhöhung der Größe und der Polarität einer Aminosäure am ω-Ende der Kette inhibiert werden kann.

Tabelle 9: Mutationen der Fettsäurebindungsstelle für die Lipasen aus *Rhizopus delemar* und *Rhizopus oryzae* und deren Effekte auf die Kettenlängenspezifität. Die Aminosäurepositionen, die der homologen RML entsprechen, sind in Klammern gesetzt.

Lipase Organismus	Mutation	Struktureller Effekt der Mutation (Abbildung 31e)	Experimentelle Beobachtung	Ref.
R.oryzae	F95Y (F94)	blockiert Bindung am C12	60% (30%) Aktivitätsanstieg für Capronsäuremethylester relativ zu Öl- (Stearin-) säure	121
R.delemar	F95D (F94)	hydrophil am C12	zweifache Aktivitätsverringerung der Hydrolyse	122
R.delemar	F95D (F94) F214R (F213)	rechte Wand des hydrophoben Spalts wird hydrophil (C10-C16)	dreifacher Aktivitätsanstieg für Tricaprylin relativ zu Triolein	123
R.delemar	F112W (F111)	blockiert Bindung am C6	50% Aktivitätsanstieg für Tributyrin relativ zu Triolein	122
R.delemar	F112Q (F111)	hydrophil am C6	keine Aktivität	
R.delemar	V206T (V205)	hydrophil am C4	10-20% Aktivität	122
R.delemar	V209W (L208)	blockiert Bindung am C10	zweifacher Aktivitätsanstieg für Tributyrin relativ to Triolein	122
R.delemar	V209W (L208) F112W (F111)	blockiert Bindung am C10 und C6	80-facher Aktivitätsanstieg für Tributyrin relativ zu Tricaprylin; keine Triolein Hydrolyse	123
R.oryzae	F214Y (F213)	blockiert Bindung am C16	20% Aktivitätsanstieg für Capronsäuremethylester relativ zu Öl- und Stearinsäure	121

4.2.3 Weiter für die Fettsäurekettenlängenspezifität entscheidende Einfüße.

Die Fettsäurebindungsstelle ist nicht der einzige Faktor, der die Kettenlängenspezifität bestimmt. Mutationen nahe des katalytisch aktiven Zentrums und des Lids können die Kettenlängenspezifität ebenfalls beeinflussen. Der Austausch der *oxyanion hole* bildenden Aminosäure T83 gegen ein Serin in *Rhizopus delemar* verringert die Aktivität gegenüber Tributyrin vierfach relativ zu Triolein.[122] Der Austausch des im Lid lokalisierten W89 in der Lipase aus *Humicola lanuginosa* gegen eine kleinere und polarere Aminosäure erhöht die Triacetin und Trioctanoin Aktivität relativ zu Tributyrin.[122] Diese Ergebnisse stehen in Einklang mit experimentellen Beobachtungen, dass die Aktivität und Spezifität von Lipasen durch mehrere Faktoren, wie Wasseraktivität,[124] Immobilisierung[125] und der Güte der Substratgrenzfläche,[126] beeinflusst werden kann. Die Mutation des katalytischen Zentrums oder des Lids könnten diese Faktoren beeinflussen und die Wechselwirkung der Lipase mit der Substratgrenzfläche ändern, falls die Aktivität an der Wasser-Ölgrenzschicht gemessen wird. Jedoch wurde gezeigt, dass experimentelle Methoden, die kompetitive Faktoren α in organischen Lösungsmittel messen, diese sekundären Effekte ausschließen.[127] Für ein Substratpaar entspricht α dem Verhältnis der Spezifitätskonstanten k_{cat}/K_m: in der Alkoholyse und der Veresterung ist α eher abhängig von der Fettsäurekette als von den Reaktionsbedingungen oder der chemischen Struktur des Alkohols.[127]

4.2.4 Fettsäurekettenlängenspezifität der RML und CALB

Die Kettenlängenspezifität der RML und CALB in der Alkoholyse und Veresterung von Monoestern wurde in organischem Lösungsmittel[127,128] anhand des kompetitiven Faktors α gemessen (Abbildung 39). Die Kettenlängenprofile der Lipasen unterscheiden sich deutlich. Unter diesen experimentellen Bedingungen ist die Fettsäurebindungstelle der bestimmende Faktor für die Kettenlängenspezifität, wogegen die Wahl der Reaktionsbedingungen eher untergeordnet ist. CALB zeigt eine hohe Aktivität für kurze und mittellange Fettsäureketten und eine geringere Aktivität für langkettige Fettsäuren. Dagegen zeigt RML eine relativ geringe Aktivität gegenüber kurzkettigen Fettsäuren, jedoch eine ansteigende Aktivität für längerkettige Fettsäuren. Dieses Verhalten steht in Einklang mit der Form und den Eigenschaften der jeweiligen Fettsäurebindungsstelle (Abbildung 30, Abbildung 31). Die Fettsäurebindungsstelle der CALB ist relativ kurz (C13) und besitz einen kleinen hydrophoben Bereich am Rand der trichterförmigen Bindungstasche. Für die RML liegt diese in einem langen (C18) deutlich ausgebildeten hydrophoben Spalt.

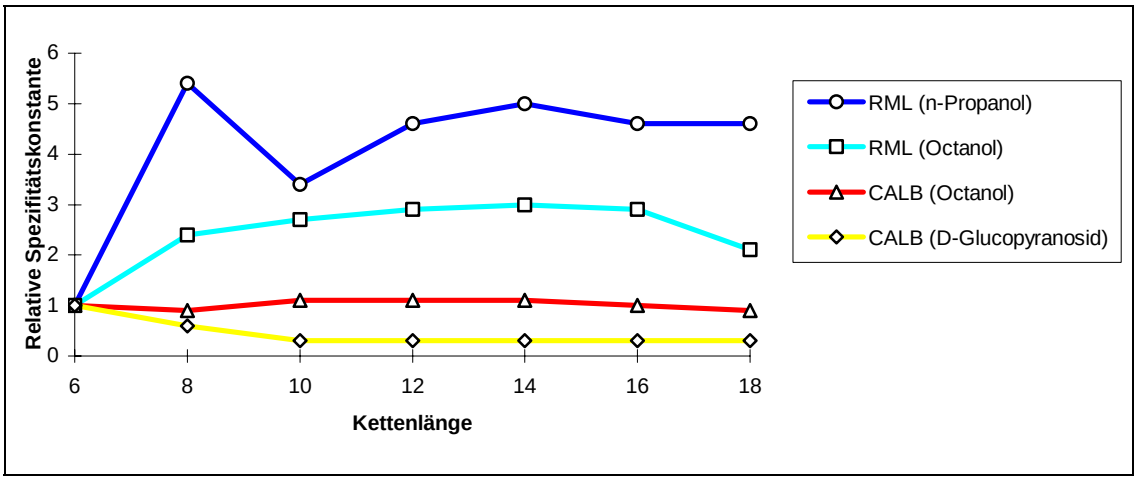


Abbildung 39: Relative Spezifitätskonstanten k_{cat}/K_m für RML und CALB (auf C6 normalisiert) gegenüber Fettsäureketten unterschiedlicher Länge. RML: Alkoholyse von Monoester mit n-Propanol[127] und Veresterung von Octanol in Hexan;[128] CALB: Veresterung von Octanol in Hexan[128] und Veresterung von D-Glucopyranosid.[128]

Wenn jedoch Assays angewandt werden, die die Aktivität an der Wasser-Substratgrenzfläche bestimmen, können drastische Änderungen der Kettenlängenspezifität auftreten, wie es für die Hydrolyse von Triglyceriden mit RML gezeigt werden konnte.[129] Zusätzlich zur Fettsäurebindungsstelle können auch andere strukturelle Elemente wie die Alkoholbindungsstelle und das Lid die Kettenlängenspezifität beeinflussen.

4.3 Funktionsanalyse

Durch die eigenständigen Analysen der Sequenz- und Strukturdaten der α/β-Hydrolasen konnten bereits neue Erkenntnisse für diese Proteinfamilie gewonnen werden. Aber erst die Kombination dieser beiden Methoden führte zu einem tieferen Verständnis funktioneller Konzepte. Die Identifizierung von lokalen Sequenzmotiven, die innerhalb der Familien konserviert waren, führte erst in Verbindung mit der strukturellen Analyse des *oxyanion holes* einzelner α/β-Hydrolasen zur Beschreibung einer substratspezifischen Signatur und der Entwicklung des Anker Konzepts, das experimentell bestätigt werden konnte. Die innerhalb von Superfamilien streng konservierten Aminosäuren wurden durch den Vergleich verschiedener Proteinstrukturen als der Faltungsnukleus identifiziert und ein konzeptionelles Modell der Faltung der α/β-Hydrolasen konnte skizziert werden.

4.3.1 Funktionsspezifische Signaturen

Während die Architektur der α/β-Hydrolasen und die Geometrie der katalytischen Triade erhalten sind, kann das *oxyanion hole* in drei Klassen eingeteilt werden: die GX, GGGX und Y Klasse, die unterschiedliche Konzepte der *oxyanion hole* Stabilisierungen aufweisen. Für die GX und GGGX Klasse, ist die Sequenz des *oxyanion holes* jedoch ein Motiv, das die Vorhersage der Struktur des *oxyanion holes* ermöglicht. Ebenso spiegeln diese Klassen Unterschiede in der Substratspezifität wieder: das GGGX Motiv tritt in Carboxylesterasen und kurzkettenspezifische Lipasen auf, das GX Motiv in Lipasen mit Substratspezifität gegenüber mittel- und langkettige Fettsäuren. Dies konnte für die GGGX Klasse experimentell gezeigt werden.[119] Für Lipasen und Carboxylesterasen sind in der Sequenzmusterdatenbank Prosite[18] zwei Muster bekannt, die das funktionell relevante und konservierte GXSXG Motiv beinhalten. Zusätzlich ist für Carboxylesterasen ein weiteres, strukturell relevantes Muster beschrieben, das den β-Strand β2 des kanonischen α/β-Hydrolase Folds umfasst. Das GGGX Motiv ist bis jetzt noch nicht als carboxylesterasespezifisches Muster identifiziert, da es hoch repetitiv in einer großen Anzahl von verschiedenen Proteinen vorkommt. Jedoch in Verbindung mit dem GXSXG Motiv kann GGGX zu einem Muster spezifisch für Carboxylesterasen kombiniert werden. Die Spezifität des Muster kann durch den zweiten Oxyanionholebildners im GXSXG Motiv, der ein gut konserviertes Alanin ist, noch verbessert werden. Des Weiteren steht X in GGGX überwiegend für F, L oder Y. Daraus ergibt sich das Muster GGG[F,L,Y]-$X_{n(n=70...100)}$-GXSAG, das hoch spezifisch für funktionelle Bereiche der Carboxylesterasen ist.

4.3.2 Das Anker Konzept

Kristallstrukturen zeigen, dass Lipasen mehrere Bereiche besitzen, die für die spezifische Funktion verantwortlich sind. Die Hydrolyse der Esterbindung wird durch die katalytische Triade bewerkstelligt, die für den nukleophilen Angriff an das Substrat verantwortlich ist. Diese wird durch das *oxyanion hole* unterstützt, das den tetraedrischen Übergangszustand, der während der Acylierung und Deacylierung des Substrats auftritt, über Wasserstoffbrücken der Backboneamidgruppen stabilisiert.[130] Es wird durch zwei Aminosäuren gebildet wovon eine der C-terminal Nachbar des katalytisch aktiven Nucleophils ist und die zweite N-terminal dazu lokalisiert ist. Für die GX Klasse stellt X die N-terminal positionierte *oxyanion hole* Aminosäure dar, die in einem Loop ohne Sequenz- und Strukturkonservierung lokalisiert ist. Jedoch ergab die Analyse der Superfamilien der LED, dass X innerhalb dieser Familien streng

konserviert ist. In der Familie abH23 wird X durch eine polare Aminosäure mit Hydroxylgruppe repräsentiert. Für einige Vertreter dieser Familie wurden die *oxyanion hole* Aminosäuren experimentell bestimmt. Für HLL stellt S83 die *oxyanion hole* bildende Aminosäure dar.[131] Yamaguchi et al. zeigten mit Hilfe der *site-directed Mutagenese*, dass S83 das *oxyanion hole* in PeCl bildet,[132] und in FHL ist S82 die entscheidende Aminosäure.[133] In RDL[131] und RNL[134] wird die N-terminale *oxyanion hole* Aminosäure durch ein Threonin gebildet (T83).

Für Lipasen der Familie abH23.1 und Cutinasen der Familie abH36.3 sind experimentelle Daten für *oxyanion hole* Mutanten bekannt. Der Austausch des T83 in RDL gegen ein Alanin eliminiert die Aktivität der Lipase gegenüber Tributyrin, Triolein und Tricaprylin.[122] T82 des *oxyanion holes* der ROL wurde gegen Valin und Alanin ausgetauscht.[114] Diese Mutationen führten zu einer Restaktivität von 0,04% bzw. 0,12%. Diese Ergebnisse unterstützen die bekannte Hypothese, dass die Hydroxylgruppen dieser Aminosäuren eine zusätzliche Stabilisierung des Übergangszustandes bewirken.[83,131] Jedoch führt der Austausch des Threonins gegen ein Serin, das der *oxyanion hole* Aminosäure in RML entspricht, ebenfalls zu einem Aktivitätsverlust. Gegenüber Triolein ergab sich eine Restaktivität von 12% ohne eine signifikante Änderung des K_M-Wertes und 14,5% Restaktivität gegenüber Butylester. Dieser Aktivitätsverlust wurde durch die geringere Tendenz des Serins Wasserstoffbrücken zum Oxyanion auszubilden erklärt. Diese Vermutung wurde aus Molekulardynamischen Simulationen abgeleitet.[114] Der Austausch der *oxyanion hole* Aminosäure S42 in FSC gegen ein Alanin führte ebenfalls zu einem drastischen Aktivitätsverlust (450-fach) ohne signifikante Änderung in der Kristallstruktur.[135]

Im Gegensatz zu diesen Hydrolasen, für die vermutet wird, dass die Hydroxylgruppe der *oxyanion hole* Aminosäure zur Stabilisierung des Übergangszustandes des Substrats beiträgt, besitzen eine Vielzahl von Hydrolasen an Stelle der polaren, eine hydrophobe Seitenkette. Für die Lipasen HuPL und PPL der Familie abH20.3 bildet L154 die N-terminale *oxyanion hole* Aminosäure[136,137] und in HPL und GPL die Aminosäure L153.[138,139] In der Familie abH15.2 wird in den Lipasen BCL,[101] BGL[140] und CVL[141] die N-terminale *oxyanion hole* Aminosäure durch L17 gebildet. In DGL als Vertreter der Familie abH14.2 wird das *oxyanion hole* durch die Aminosäure L67 gebildet.[142] Somit können diese *oxyanion hole* Aminosäuren nicht durch eine zusätzliche Wasserstoffbrücke zur Stabilisierung des Oxyanion beitragen.

Die systematische Analyse der Kristallstrukturen für Hydrolasen der GX Klasse in der offenen Konformation ergab, dass die Seitenkette der N-terminalen *oxyanion hole* Aminosäure mit mindestens einer weiteren Aminosäure interagiert. Im Gegensatz zur

Hypothese der zusätzlichen Stabilisierung des Übergangszustandes für die Lipasen der Familie abH23.1, führen die hier beschriebenen Ergebnisse zu der Hypothese, dass die Interaktion der Seitenkette mit dem sogenannten Anker für die Stabilisierung der *oxyanion hole* Geometrie entscheidend ist und somit wichtig für die enzymatische Aktivität. Um diese Hypothese zu validieren, wurden das hydrophobe Anker-*oxyanion hole* Modul der BCL gegen das hydrophile Modul der RML schrittweise ausgetauscht. Der Austausch der *oxyanion hole* Aminosäure L17 gegen ein Threonin resultierten in einem drastischen Aktivitätsverlust gegenüber pNPP und Olivenöl. Ebenso führt der Austausch gegen ein Serin oder Glycin zu einem hohen Aktivitätsverlust (persönliche Mitteilung Dinh Thi Quyen). Diese Ergebnisse deuten darauf hin, dass die fehlende Stabilisierung durch die Interaktion mit dem Anker zu einer Deplatzierung des Loops führt, der die *oxyanion hole* Aminosäure trägt.

Die strukturelle Nähe des Ankers zur *oxyanion hole* Aminosäure sowie die experimentellen Befunde für die Einfachmutanten der BCL lassen eine funktionelle Wechselwirkung dieser Aminosäuren vermuten. Die Austauschbarkeit dieses Moduls wurde bisher jedoch noch nicht nachgewiesen. Durch den Austausch des Ankers L167 in der inaktiven Variante der BCL mit der *oxyanion hole* Mutation L17T gegen die Anker Aminosäuren Aspartat, Glutamat, Asparagin und Glutamin, die in Familie abH23.1 auftreten, führt zu einer Reaktivierung der BCL Einfachmutante. Bis auf die Doppelmutante L17T/L167E, die in keiner aktiven Form erhalten werden konnte, konnte so eine Aktivitätssteigerung um eine Größenordnung erreicht werden. Jedoch scheint die Interaktion zwischen Anker und *oxyanion hole* Aminosäure ein durch die Evolution hoch optimiertes System darzustellen und reagiert somit sensitiv auf Störungen, weshalb die Zurückgewinnung der Gesamtaktivität des Wildtyps nicht erreicht werden konnte. Dies wird ebenso durch den Austausch des Ankers D91 in ROL gegen ein weniger polares Asparagin deutlich. Dieser Austausch schwächt die Interaktion zwischen Anker und *oxyanion hole* Aminosäure und führt so zu einer Verringerung der Aktivität.[114]

Die Identifikation dieses funktionell relevanten Konzepts spielt auch für das *Protein-Engineering* eine entscheidende Rolle, da der Anker in vielen Fällen Teil der Substratbindungstasche ist. Yang et al. untersuchten die Kettenlängenspezifität der Lipase aus *Burkholderia cepacia* KWI-56, die 94% Sequenzidentität mit BCL besitzt, über kombinatorische Mutagenese.[143] Da der Anker L167 am Ende des hydrophoben Spalts der Substratbindungstasche lokalisiert ist, in dem Substrate bis zum sechsten Kohlenstoffatom binden, wurde der Einfluss dieser Position auf die Kettenlängenspezifität untersucht. Eine drastische Änderung der Kettenlängenspezifität wurde durch die Mutation L167V erreicht. Eine 1,5-fache Verringerung der spezifische Aktivität gegenüber p-Nitrophenylcaprylat sowie

eine 2,1-fache Verringerung gegenüber p-Nitrophyenlpalmitat wurde beobachtet. Die spezifische Aktivität gegenüber p-Nitrophyenlbutyrat wurde nicht geändert. Diese Effekte können als Ergebnis einer Kombination zweier Einflüsse erklärt werden. Während die unterschiedliche Verringerung der Aktivität für verschiedene Substratlängen eine Folge der geänderten Bindungsbedingungen ist, ist eine Erhöhung der Aktivität aufgrund der sensitiven *oxyanion hole*, Anker Interaktionen nicht zu erwarten. Koga et al. wählten ebenfalls den Anker L167, die *oxyanion hole* Aminosäure L17 sowie weitere Positionen der Bindungstasche in BCL KWI-56 für die Untersuchung der Enantioselektivität durch kombinatorische Mutagenese.[144] Varianten wurden von Koga et al. beschrieben mit Aktivität gegenüber p-Nitrophenyl-3-phyenlbutyrat trotz des Austauschs des Ankers und der *oxyanion hole* Aminosäure gegen ein Glycin oder Alanin. Diese Varianten sollten aufgrund der fehlenden *oxyanion hole* Stabilisierung durch die Interaktion des Ankers mit der *oxyanion hole* Aminosäure keine Aktivität aufweisen. Jedoch wurden Aktivitäten nur für Varianten gemessen, die die Mutationen L17F/L167G oder L17F/L167A enthielten. Diese Varianten waren von besonderem Interesse, da sie gegenüber den untersuchten Substraten eine umgekehrte Enantioselektivität aufwiesen. Die L17F/L167G enthaltenden Varianten zeigten sogar eine 1,5-fach erhöhte Aktivität jedoch mit einer geringern spezifischen Aktivität (90,6 U/mg). Koga et al. erklärten die umgekehrte Enantioselektivität dadurch, dass L167G die Bindungstasche für die sehr sperrige Phenylbutyratgruppe vergrößert. Somit würde die Phenylbutyratgruppe den Bereich zwischen dem Anker und der *oxyanion hole* Aminosäure besetzen und könnte somit die Rolle des Ankers übernehmen und zur Stabilisierung des *oxyanion holes* beitragen. Dies würde eine spezielle Form der substratunterstützten Katalyse darstellen, wie für verschiedene Enzymklassen bereits beschrieben.[145] Während in den meisten Fällen das Substrat eine funktionelle Gruppe ersetzt, die direkt an der chemischen Reaktion beteiligt ist, übernimmt hier das Substrat die Rolle einer für die Funktion indirekt beteiligten Gruppe, die für die Orientierung des katalytischen Apparates notwendig ist. Die ebenfalls eine umgekehrte Enantioselektivität aufweisende Variante mit L17F/L167A zeigt eine 5,5-fach geringere Aktivität im Vergleich zur L17F/L167G Variante. Da die Methylgruppe des Alanins den Raum zwischen Anker und *oxyanion hole* Aminosäure verkleinert, könnte das *oxyanion hole* durch die sterisch anspruchsvolle Phenylbutyratgruppe stärkere deplaziert sein und so zur verringerten Aktivität führen.

Für hydrophile Module scheint nicht nur die direkte Stabilisierung des *oxyanion holes* entscheidend zu sein, sondern auch das Wasserstoffbrückennetzwerk in das die Module involviert sein können. Herrgard et al. untersuchten das elektrostatische Netzwerk für

Mitglieder der Familie abH23.[146] Dieses Netzwerk enthält in RML neben der katalytischen Triade und geladenen Aminosäuren im Lid auch den Anker D91. Die theoretische Studie untersuchte pKa Verschiebungen für verschiedene Varianten der RML. Für D91N konnte gezeigt werden, dass diese Mutation in der offen Konformation einen Einfluss auf den pKa des katalytisch aktiven Histidins hat und so die Aktivität des Enzyms beeinflussen kann. Somit müssen solche Effekte beim Austausch eines hydrophilen gegen ein hydrophobes Modul ebenfalls berücksichtigt werden.

4.3.3 Analyse der positionsspezifischen Konservierung

Die Bestimmung der Aminosäurekonservierung innerhalb von Multisequenz Alignments für Proteinfamilien wurde bereits erfolgreich für die Untersuchung der Proteinfaltung eingesetzt. Friedberg und Mergalit analysierten die Konservierung von Aminosäuren in strukturell ähnlichen aber in der Sequenz unterschiedlichen Proteinen.[147] Sie entdeckten Aminosäurepositionen, die innerhalb einer Familie konserviert sind, sich aber zwischen verwandten Familien unterscheiden. Mirny und Shakhnovich beobachteten das gleiche Phänomen und bezeichneten es als „Conservation of Conservation“ (CoC).[46] Sie folgerten, dass diese Motive ein Ergebnis des evolutiven Drucks auf die Protein Faltung sind. Bei der Analyse der Aminosäure Konservierung sind jedoch zwei Probleme zu beachten. Die verwendeten Multisequenz Alignments müssen korrekt sein, da schlechte Alignments biologisch relevante Konservierungen verschleiern können. Es muss sicher sein, dass eine konservierte Position aufgrund der Funktion im Alignment erhalten ist und nicht Mangels einer zu geringen Anzahl oder zu geringen Diversität der verwendeten Sequenzen. Des Weiteren muss das Strukturalignment, das als Schnittstelle zwischen den Multisequenz Alignments der Superfamilien dient verlässlich sein.

Die Qualität der verwendeten Multisequenz Alignments wird über die Klassifikation der LED gewährleistet. Da diese Klassifikation im Unterschied zu voll automatisierten Klassifikationsansätzen auf die Erhaltung für die Funktion wichtiger Aminosäuren basiert sind familienfremde Einträge unwahrscheinlich. Des Weiteren wurden Fragmente nicht berücksichtig.

Die zweite Frage ist ob beobachtete Konservierungen real sind oder nicht. In dieser Studie wurden nur Superfamilien in Betracht gezogen die mindestens 20 Sequenzen mit einer Sequenzidentität weniger als 90% besitzen. Somit ist die Möglichkeit gegeben, dass alle Aminosäuren in einer Spalte auftreten können. Tatsächlich beinhalteten alle untersuchten Superfamilie mit einer Ausnahme deutlich mehr als 20 Sequenzen, so dass die Diversität hoch genug für eine verlässliche Analyse war.

Für die Bestimmung der streng konservierten Positionen, die zwischen den Familien korrelieren, ist das Strukturalignment entscheidend. Für α/β-Hydrolasen zeigte sich hierbei die Schwierigkeit, dass entscheidende Elemente des α/β-Hydrolase Fold stark variieren können. So kann die Länge des β-Strands β5 zwischen 5 und 13 Aminosäuren, die Länge der α-Helix αE zwischen 4 und 17 Aminosäuren betragen. Des Weiteren ist der Grad der Verdrillung des zentralen β-Faltblatts für verschiedene α/β-Hydrolasen unterschiedlich, so dass eine Überlagerung nur über dieses strukturelle Merkmal zur Deplatzierung der peripheren Sekundärstrukturelemente führt. Durch die für diese Analyse jedoch optimierte Methode der Überlagerung und das manuelle Überarbeiten führte zu einem verlässlichen Alignment. So stellen die 106 überlagerten Positionen des strukturellen Alignments 42% der Referenzstruktur 1J2E der Familie abH27 dar, wobei diese α/β-Hydrolase mit 251 Aminosäuren eine der kürzeren Vertreter dieses Proteinfaltungsmusters darstellt. Bezieht man die 106 überlagerten Positionen auf die Aminosäuren der 1J2E, die Teil der für den kanonischen α/β-Hydrolase Folds elementaren Sekundärstrukturelemente sind, ergibt sich eine Abdeckung von 88%. Somit ist der Grossteil eines minimalen α/β-Hydrolase Folds durch das strukturelle Alignment repräsentiert.

4.3.4 Der Faltungsnukleus der α/β-Hydrolasen

Die Familie der α/β-Hydrolasen stellt eine der größten bekannten Proteinfamilien dar, die einen hoch diversen Sequenzraum aufspannen mit Sequenzidentitäten unter 10%. Dennoch weisen alle Proteinstrukturen der α/β-Hydrolasen das selbe Faltungsmuster auf. Allgemein zeigt die Verteilung der bekannten Proteinstrukturen auf die in CATH definierten Folds ein deutliches Übergeweicht für eine kleine Anzahl an Folds auf, die etwa ein Drittel aller in der PDB abgelegten Proteinstrukturen in sich vereinigen. Salem et al. zeigten, dass in diesen sogenannten Superfolds[4] die Häufigkeit verschiedener Sekundärstrukturmuster signifikant erhöht ist, im Vergleich zu nicht Superfolds.[8] Diese Muster, auch Supersekundärstrukturen (SSS) genannt, sind der β- und α-*Hairpin*, die βαβ-Einheit sowie die etwas komplexeren Einheiten β_4- und βα-*Greek key*. In den Superfolds stellen diese SSS über 60% aller Aminosäuren, die in der Ausbildung von Sekundärstrukturelementen beteiligt sind. Diese Elemente könnten eine entscheidende Rolle für die Faltung der Proteine spielen. Da diese Elemente energetisch günstige Zustände mit einer bevorzugten Konfiguration darstellen, ist es wahrscheinlich, dass an diesen Positionen die Faltung initialisiert wird. Da die Superfolds eine Vielzahl dieser SSS enthalten, sind diese kinetisch begünstigt, da die Faltung an

beliebigen Stellen in der Sequenz beginnen kann. Diese lokale Initialisierungselemente der Faltung entsprechen somit den von Fersht postulierten Foldons.[148]

Auch für die α/β-Hydrolasen sind solche SSS in der PDBsum[99] beschrieben. Für dieses Faltungsmuster finden sich überwiegend βαβ-Elemente, die den Kern der α/β-Hydrolasen stellen. So sind nach der Definition durch PROMOTIF[149] für den kanonischen α/β-Hydrolase Fold vier βαβ SSS zu erwarten: β3αAβ4, β5αCβ6, β6αDβ7 und β7αEβ8. Für die Referenzstrukturen der acht untersuchten Superfamilien zeigt sich mit zwei Ausnahmen, dass diese vier βαβ Elemente vorhanden sind oder wie im Fall der Familie abH4 mit erweitertem kanonischen α/β-Hydrolase Fold um ein weiteres βαβ Element ergänzt ist. Für die Referenzstruktur 1K8Q der Familie abH14 waren nur drei und für die Referenzstruktur 4LIP der Familie abH15 sogar nur ein βαβ Element vorhanden. Dies liegt in der Einlagerung von weiteren kurzen β-Strands innerhalb dieser Elemente begründet. Dies steht aber in Übereinstimmung mit Untersuchungen anderer Faltungsmuster des *doubly wound* Folds.[8] Dieser Superfold weist allgemein einen geringeren Prozentsatz an SSS im Vergleich zu anderen Superfolds auf. Auch in diesen Fällen sind Sekundärstrukturelemente, die zwischen den β-Strands eines potentiellen βαβ Elements eingelagert sind, die Ursache. Das Element β4αBβ5 ist aufgrund der Topologie nicht als βαβ Element definiert.

Bildet man die Positionen der streng konservierten Positionen auf diese SSS Elemente ab, zeigt sich, dass diese nicht gleich verteilt sind. Die C-terminal lokalisierten βαβ Elemente β6αDβ7 und β7αEβ8 besitzen deutlich weniger SKPs. Diese sind zudem in dem Bereich des α/β-Hydrolase Folds lokalisiert, der die größte Variabilität aufweist. Zwischen den β-Strands β6 und β7 sowie zwischen β7 und β8 finden sich große Insertionen wie das Lid oder Cap Elemente, die sich in ihrer Länge und in der Aminosäurekomposition stark unterscheiden. Im Gegensatz zu diesem Bereich, enthalten die N-terminal lokalisierten βαβ Elemente β3αAβ4 und β5αCβ6 den Grossteil der SKPs, die überwiegend hydrophob und für das Lösungsmittel nicht zugänglich sind. Auffällig ist, dass die α-Helix αA, die Teil des βαβ Elements β3αAβ4 ist, keine konservierten Positionen aufweist. Die α-Helix αB, die zwischen β3αAβ4 und β5αCβ6 lokalisiert und nicht Teil eines SSS Elements ist, enthält jedoch den größten SKP Anteil aller α-Helices. Betrachtet man die räumliche Lage der α-Helices, so zeigt sich, dass die α-Helix αA durch das zentrale β-Faltblatt von den α-Helices αB und αC separiert ist. In Verbindung mit den oszillierenden SKPs der β-Strands bilden die konservierten Seitenketten der β-Strands β3 und β4, der α-Helix αB sowie des SSS Element β5αCβ6 den kompakten Kern der α/β-Hydrolasen. Eine solche Clusterbildung von dicht gepackten,

konservierten Aminosäuren wurden auch von Mirny und Shakhnovich beobachtet. In ihrer Untersuchung der fünf häufigsten Folds konnten sie zeigen, dass Aminosäuren mit hoher CoC solche Cluster bilden, diese sich aber in der Lage und der Art der Wechselwirkungen zwischen den Folds unterscheiden können.[46]

Während die SSS Elemente durch ihre energetisch günstige Geometrie die Faltung lokal initialisieren, werden die von Mirny und Shakhnovich beschriebenen in der Sequenz nicht lokalen Cluster von dicht gepackten, konservierten Aminosäuren als spezifische Faltungsnuklei bezeichnet.[150,151] Dieses Faltungsnukleusszenario geht davon aus, dass nachdem der spezifische Nukleus ausgebildet wurde, die Proteinfaltung nachfolgende Übergangszustände schnell durchläuft, bis die native Proteinstruktur erreicht ist.

Für die Faltung des spezifischen Nukleus der α/β-Hydrolasen lässt sich somit folgender Verlauf skizzieren. Nach der lokalen Ausbildung der βαβ Elemente β3αAβ4 und β5αCβ6, die über α-Helix αB miteinander verbunden sind, müssen sich diese beiden Elemente um αB positionieren um den dicht gepackten Nukleus zu bilden. Dieser Schritt wird sicherlich durch den hydrophoben Kollaps bewirkt. Auffällig sind jedoch die streng konservierten Positionen SKP_{16} und SKP_{19}. Diese bilden in den meisten untersuchten Fällen eine Salzbrücke aus oder wechselwirken über ihre hydrophilen Seitenketten. SKP_{16} ist am C-terminalen Ende des β-Strands β4 und SKP_{19} am N-terminalen Ende der α-Helix αB lokalisiert. Die Lage dieser Aminosäuren, sowie die weitreichende, attraktive Wechselwirkung könnten die Anlagerung des SSS Elements β3αAβ4 an die α-Helix αB bewirken: nach der Ausbildung dieser beiden Elemente müssen sich diese noch nicht räumlich nahe stehen. Durch die weitreichende Wechselwirkung zwischen SKP_{16} und SKP_{19} könnten diese beiden Elemente sich gegenseitig einfangen und so deren Zusammenballung initialisieren. Nach der Ausbildung des Nukleus bilden diese beiden Positionen eine zusätzliche Stabilisierung des Clusters aus.

Die Topologie der Lipase aus *Rhizomucor miehei* und deren Homologen unterscheidet sich von der des kanonischen α/β-Hydrolase Folds in der Anordnung der β-Strands des zentralen β-Faltblatts. Diese Lipasen scheinen im Laufe der Evolution den β-Strand β4 verloren zu haben und besitzen somit das SSS Element β3αAβ4 nicht. Dies führt auch zum Verlust des SKP_{16}. Die Position SKP_{19} ist jedoch innerhalb der Familie durch ein gut konserviertes Glutamat besetzt. Die Strukturanalyse der RML zeigt, dass diese Position weiterhin zur Stabilisierung des Faltungsnukleus beiträgt. Die Wechselwirkung mit der Position SKP_{16} ist in dieser Familie durch drei Wasserstoffbrücken mit den Amidgruppen des Backbones der Aminosäuren L58, I59 und Y60, die den Loop zwischen β-Strand β1 und β2 bilden, ersetzt.

Dies ist ein weiteres Indiz für die Relevanz der Positionen SKP_{16} und SKP_{19} in der Faltung der α/β-Hydrolasen.

5 Material und Methoden

5.1 Betriebssystem

Das Data Warehouse System wurde auf einer Dual Athlon MP Plattform unter RedHat Linux 9.0 entwickelt. Die Entwicklungsumgebung wurde unter Microsoft Windows 2000 auf einer Athlon XP Workstation eingesetzt. Die Berechnung der Maximum Likelihood Bäume wurde auf einem Athlon MP/Mandrake Linux 9.1 Cluster mit 128 Knoten durchgeführt. Die Berechnung der Lösungsmittelzugänglichkeit mit NACCESS sowie die Berechnung der GRID Bindungspotentiale wurden unter IRIX 6.5 auf einer SGI Octane 2 MIPS 12000 Workstation durchgeführt.

5.2 Relationales Datenbank Management System (RDBMS)

Verwendet wurde das Open Source Datenbank System Firebird 1.5, wobei die SuperServer Engine eingesetzt wurde. Firebird basiert auf dem von Borland Software Corp veröffentlichten InterBase 6.0 Quellcode unter der InterBase Public License V.1.0 vom 25 Juli 2000. Firebird ist unter http://www.ibphoenix.com erhältlich.

Da die Entwicklung des Data Warehouse Systems möglichst plattformunabhängig sein sollte und für ein nicht kommerzielles, universitäres Umfeld entwickelt wurde, fiel die Wahl auf ein freies Open Source RDBMS. Aus den folgenden Gründen wurde nicht MySQL als populärster Vertreter freier RDBMS, sondern Firebird gewählt.

Firebird kann mit InterBase als Vorläufer, dessen Entwicklung 1985 begann, auf fast 20 Jahre Datenbankerfahrung zurückblicken. Diese lange Entwicklungszeit und der langjährige Einsatz in verschiedenen kommerziellen Umgebungen spricht für ein ausgereiftes Produkt. Firebird unterstützt vollständig den SQL-92 Standard und implementiert einen Großteil des SQL-99 Standards. Im Vergleich zu MySQL bietet Firebird sogenannte Stored Procedures. Diese sind in die Datenbank eingebettet und erlauben so das serverseitige Verarbeiten von Daten. Dies ermöglicht das Auslagern von sich wiederholenden Aufgaben aus der Applikation in die Datenbank. Firebird stellt Trigger bereit die vor oder nach *Insert*, *Update* oder *Delete* Ereignissen ausgelöst werden können. In Verbindung mit den implementierten Generators lassen sich so selbstinkrementierende Felder erzeugen. Des Weiteren bietet Firebird ein online Backup mit dessen Hilfe das Portieren einer Datenbank auf andere Plattformen möglich ist.

Für den Remotezugang wurde das Open Source Administrationswerkzeug IBOConsole Version 1.1.9.6 verwendet. Diese basiert auf der von Borland veröffentlichte IBConsole.

5.3 Programmiersprache der Anwendungsentwicklung

Die Entwicklung des Data Warehouse Systems wurde vollständig in Perl (Practical Extraction and Report Language) Version 5.8.0 realisiert. Perl ist als Open Source unter der Perl artistic Lizenz veröffentlicht und ist über http://www.perl.com/download.csp zugänglich. Ein Vielzahl weiterer Perl Module und ergänzende Informationen sind über das Comprehensive Perl Archive Network (CPAN) unter http://www.perl.com/CPAN/ erhältlich.
Perl ist plattformunabhängig und besitzt umfangreiche Funktionen zur Mustererkennung, die das Verarbeiten von Text wesentlich vereinfachen. Des Weiteren erlaubt Perl das einfache Einbinden externer Programme und über das Common Gateway Interface (CGI) eine webserverseitige Implementierung. Für alle gängigen RDBMS stehen Treiber zur Anbindung zur Verfügung.

5.4 Weboberfläche

Als graphisches Interface für das Data Warehouse System wurde eine Weboberfläche entwickelt. Dies ermöglicht den plattformunabhängigen Zugang zum System über jeden JavaScript fähigen Internetbrowser.
Statische Webseiten wurden in HTML 4.01 geschrieben. Dynamische Effekte innerhalb der Webseiten wurden mit JavaScript erzeugt.
Dynamisch erzeugte Webseiten wurden über CGI Perlskripte erzeugt, die Daten aus der Datenbank auslesen, prozessieren und in HTML formatiert im Internetbrowser ausgeben.

5.5 Editoren

Der Perl Quellcode wurde editiert und getestet mit der integrierten Entwicklungsumgebung (IDE) Komodo Version 2.5 von Activestate, a Division of Sophos, #400-500 Granville Street, Vancouver, BC, V6C 1W6 Canada http://www.activestate.com/. Allgemeines Texteditieren sowie das Erstellen von HTML Quellcode erfolgte mit Ultraedit32 Version 8.20 von IDM Computer Solutions, Inc., 3987 Hamilton Middletown Rd. Unit G, Indian Springs, OH 45011, USA http://www.ultraedit.com/.

5.6 Webserver

Für die Präsentation der Weboberfläche als auch der öffentlich zugänglichen LED wurde der Apache HTTP Server der Apache Software Foundation in der Version 2.0.46 unter Linux eingesetzt. Das Softwarepaket ist erhältlich unter http://httpd.apache.org/.

5.7 Methoden der Bioinformatik

5.7.1 Phylogenetische Analysen

5.7.1.1 Unweighted Pair Group Method using Arithmetic mean (UPGMA)

Für die Berechnung von phylogenetischen Bäumen nach der UPGMA Methode wurde der in dem Programm neighbor des Phylip Softwarepackets[10] implementierte Algorithmus verwendet. UPGMA ist eine auf Distanzen basierte Methode zur Berechnung von Phylogenetischen Bäumen und benötigt Distanzen zwischen Sequenzpaaren. UPGMA erzeugt zuerst für jede Sequenz einen Knoten und gruppiert sodann die zwei Knoten (u, v) der beiden ähnlichsten Sequenzen. So wird ein neuer Knoten (w) erzeugt mit u und v als Kindknoten und die Distanzen dieses neuen Knotens w zu allen anderen Knoten (x) wird berechnet. Dieses paarweise Gruppieren wird erschöpfend wiederholt. Die Berechnung der Distanz $D_{w,x}$ zwischen dem Knoten w und einem Knoten x erfolgt nach der Formel

$$D_{w,x} = \frac{m_u D_{u,x} + m_v D_{v,x}}{m_u + m_v}$$

wobei m_u der Anzahl an original Sequenzen im Subbaum mit Wurzel u entspricht. Das heisst, dass die Distanzen zwischen den original Sequenzen zu gleichen Teilen beitragen, weshalb die Methode auch als unweighted bezeichnet wird.

5.7.1.2 Neighbour-Joining Methode (NJ)

Für die Berechnung von Phylogenetischen Bäumen nach der NJ Methode wurde der in CLUSTALW[95] implementierte Algorithmus verwendet. Die Neighbor-Joining Methode ist wie UPGMA ein Distanz basierter Ansatz zur Berechnung von Phylogenetischen Bäumen. NJ setzt keine konstante evolutive Geschwindigkeit voraus, so dass die Theorie der Molecular Clock erfüllt sein muss. Im Gegensatz zur UPGMA Methode, die mit m Bäumen für jeden einzelnen Knoten beginnt und sukzessive den Baum vergrößert, startet die NJ Methode mit einem Baum mit minimaler Anzahl an Kanten (Stern-Topologie). Für n Sequenzen enthält dieser Baum somit einen Knoten X des Grades n (Abbildung 40a). Der Grad des Knotens X wird dann durch das Einfügen von internen Knoten sukzessive verringert, bis der Knoten X schließlich den Grad 3 erreicht. Hierfür wird jeweils ein Paar von Nachbarsequenzen ausgewählt und ein neuer Knoten Y wird erstellt mit Verbindungen zum Knoten X und den Sequenzen des Nachbarpaars. Für n Sequenzen stehen somit im ersten Schritt $n(n\text{-}1)/2$ mögliche Paare von Nachbarsequenzen zur Auswahl. Für jeden dieser Bäume wird die Summe aller Kanten berechnet und der Baum mit der geringsten Kantensumme wird

selektiert. Die Berechnung der Summe aller Kanten erfolgt z.B. für den Baum in Abbildung 40b nach der Formel

$$S_{AB} = L_{XY} + (L_{AY} + L_{BY}) + \sum_{i=C,D,E}^{N} L_{iX}$$

Da die L Werte Distanzen zwischen internen Knoten, oder internen Knoten und Blättern repräsentieren, müssen diese L Werte in der Berechnung der Summe aller Kanten in Ausdrücke umgewandelt werden, die nur die über die Distanzmatrize zugänglichen Distanzwerte enthalten. Nach dieser Transformation ergibt sich die Formel

$$S_{AB} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \le i < j} D_{ij}$$

die zur Bestimmung der Nachbarpaare genutzt wird.[152]

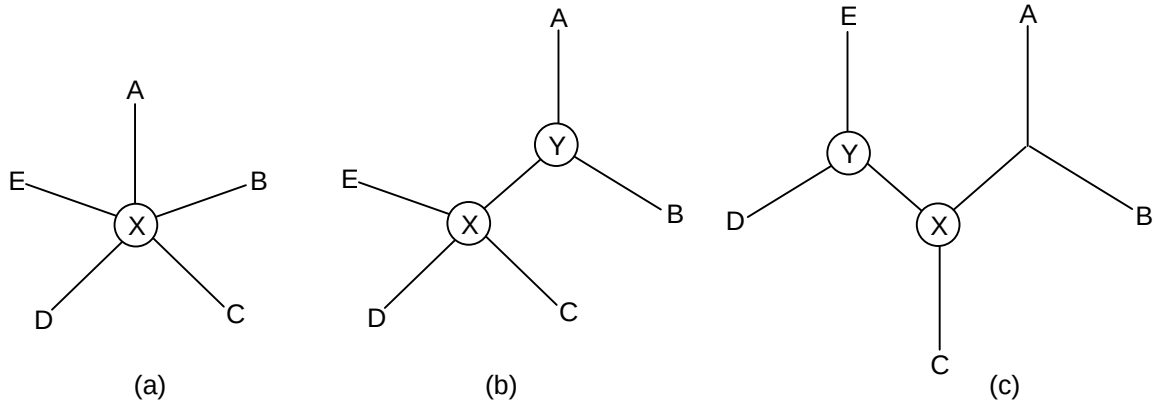


Abbildung 40: Illustration der NJ Methode. (a) Für fünf gegebene Sequenzen wird ein Baum mit Stern-Topologie konstruiert. (b) Der Grad des Knoten X wird durch Einfügen von internen Knoten Y sukzessive reduziert. (c) Die Iteration wird beendet sobald X den Grad 3 erreicht.

5.7.1.3 Maximum Likelihood Methode (ML)

Für die Berechnung von Phylogenetischen Bäumen nach der Maximum Liklihood Methode wurde der in TREE-PUZZLE 5.0[96,153] implementierte Algorithmus verwendet. ML ist eine diskrete Methode, die direkt mit Sequenzen, oder Funktionen, die aus Sequenzen abgeleitet wurden, zur Berechnung von phylogenetischen Bäumen arbeitet und nicht mit paarweise erzeugten Distanzen. ML wählt den Baum, der von allen möglichen Bäumen der wahrscheinlichste ist, um die beobachteten Daten zu erzeugen. Dieser charakterbasierte

Ansatz verfolgt somit das Ziel die Wahrscheinlichkeit zu berechnen, dass ein Satz von Daten einen gegeben Baum ergibt. Die Daten entsprechen hierbei einem Satz von n Sequenzen x^j mit $j = 1...n$ oder kurz geschrieben $x^\bullet$. Bezeichnet T den Baum mit n Blättern mit Sequenz j an Blatt j und $t_\bullet$ entspricht den Kanten des Baums, so lässt sich die Wahrscheinlichkeit als

$$P(x^\bullet \mid T, t_\bullet)$$

definieren. Um diese Wahrscheinlichkeit zu Berechnen benötigen wir ein Modell der Evolution, das z.B. Mutations- und Selektionsereignisse entlang der Kanten des Baums beschreibt.

Nehmen wir an, wir können eine Wahrscheinlichkeit $P(x|y,t)$ definieren, dass eine Vorläufersequenz y entlang einer Kante mit Länge t in eine Sequenz x übergeht. Daraus folgend kann die Wahrscheinlichkeit für einen Baum T mit einem spezifischen Satz an Vorläufern, die dessen Knoten zugewiesen werden, durch Multiplikation aller evolutiven Wahrscheinlichkeiten der Kanten berechnet werden. Z.B. für den Baum in Abbildung 41 wäre die Wahrscheinlichkeit

$$P(x^1,...,x^5 \mid T, t_\bullet) = P(x^1 \mid x^4, t_1)P(x^2 \mid x^4, t_2)P(x^3 \mid x^5, t_3)P(x^4 \mid x^5, t_4)P(x^5)$$

wobei $P(x^5)$ die Wahrscheinlichkeit beschreibt, dass x^5 die Wurzel des Baums stellt. Im Allgemeinen sind die Vorläufersequenzen unbekannt und um die Wahrscheinlichkeit $P(x^1,...,x^3|T,t_\bullet)$ der bekannten Sequenzen für einen gegeben Baum zu erhalten muss über alle möglichen Vorläufer x^4, x^5 aufsummiert werden.

Für die oben beschriebene Wahrscheinlichkeit $P(x|y,t)$ muss nun noch das evolutive Modell spezifiziert werden. Für Aminosäuresequenzen leitet TREE-PUZZLE dieses aus fixen Matrizen ab, die den Übergang einer Aminosäure in eine andere beschreiben. Standardmäßig verwendet TREE-PUZZLE hierzu die WAG Matrize, die aus einer Datenbank von 3905 Proteinsequenzen abgeleitet ist, die 182 eindeutige Familien bilden und ein breites Spektrum an evolutiven Distanzen umfassen.[154]

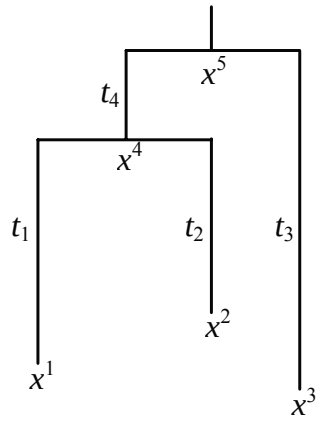


Abbildung 41: Beispiel für einen Baum drei Sequenzen.

Mit diesem Modell kann nun der ML Baum mit der Topologie T und den Kantenlängen $t_\bullet$ für den $P(x^\bullet|T,t_\bullet)$ maximal ist, gesucht werden. Diesen Baum zu finden setzt dabei voraus, dass alle möglichen Topologien für $x^\bullet$ durchsucht werden, sowie dass für jede Topologie eine Suche für alle möglichen Kantenlängen $t_\bullet$ durchgeführt wird. Da für n Sequenzen $(2n\text{-}3)!!$ Topologien möglich sind steigt der Rechenaufwand für mehrere Sequenzen somit rapide an. Um den Rechenaufwand zu minimieren hat TREE-PUZZLE deshalb eine Subbaummethode implementiert, die große Bäume aus Kleineren zusammensetzt, das Quartet Puzzling. Diese Methode besteht im wesentlichen aus drei Schritten: (1) Berechnung aller möglichen Quartet ML Bäume und Bestimmung der Quartete mit optimaler Topologie, (2) wiederholte Konstruktion (Puzzling) vollständiger Bäume der Quartet Bäume mit optimaler Topologie und (3) Bestimmung des Consensusbaums aus allen in Schritt 2 erzeugten Bäumen über eine Mehrheitsregel.[155]

Im ersten Schritt werden alle möglichen $\binom{n}{4}$ Quartette und deren ML Werte bestimmt. Für jedes Quartet (A, B, C, D) existieren 3 Topologien Q_1, Q_2 und Q_3 mit den entsprechenden ML Werten m_1, m_2 und m_3. Alle Topologien Q_i mit $m_i = \max\{m_1, m_2, m_3\}$ stellen optimale Topologien dar und werden für den Puzzling Schritt gespeichert. Für den Fall, dass mehr als eine optimale Topologie existiert, wird per Zufall ausgewählt.

Im Puzzling Schritt werden die ausgewählten Quartetbäume zu einem n-Taxonbaum kombiniert. Hierzu wird die Reihenfolge der Sequenzen zufällig gewählt (z.B. A, B, C, D, E, ...) und der optimale Baum für das Quartet (A, B, C, D) wird als Ausgang gewählt um den n-Taxonbaum zu konstruieren. Diesem Subbaum wird die Sequenz E nach folgender Auswahlregel hinzugefügt: Für jedes Quartet (i, j, k, E) wird aus dem entsprechenden optimalen ML Quartet Baum die Nachbarschaftsbeziehung abgeleitet, die aussagt, dass z.B. i und j sowie k und E über zwei interne Konten voneinander getrennt sind und somit E nicht auf einem Ast platziert werden soll, der auf dem Pfad von i nach j liegt. Die Kanten an die E

im Subbaum nicht angeknüpft werden soll werden mit einem Strafpunkt versehen. Wurden alle Quartette (*i*, *j*, *k*, E) des aktuellen Subbaums untersucht, wird E an dem Ast platziert, der die wenigsten Strafpunkte erhalten hat. Stehen mehrere Äste mit gleicher Punktzahl zur Auswahl wird aus diesen per Zufall gewählt. Da das sequenzielle Einfügen der Sequenzen entlang der zufällig generierte Ausgangsreihenfolge nicht immer zur selben Topologie führen muss, wird der Puzzlingschritt so oft wie möglich wiederholt, um den Topologieraum so gut wie möglich zu erfassen.

Im dritten Schritt wird nach einer Mehrheitsregel[155] der Consensusbaum für die im Puzzling Schritt erzeugten Bäume ermittelt. Gruppierungen von Sequenzen, die in mindestens 50% der im Puzzling Schritt erzeugten Bäume vorkommen, werden hierbei auf jeden Fall im Consensusbaum vertreten sein.

5.7.2 Globale Multisequenz Alignments

Für die Erzeugung von globalen Multisequenz Alignments wurde die in CLUSTALW 1.83 implementierte Methode verwendet.[95] CLUSTALW verwendet die heuristische Methode der progressiven Alignment Erzeugung. Dieser Ansatz basiert auf der Annahme, dass ein Alignment als eine verallgemeinerte Sequenz angesehen werden kann und so eine Sequenz mit einem Alignment oder zwei Alignments miteinander direkt verglichen werden können. Das globale Multisequenz Alignment wird so schrittweise durch paarweises Alignen von einfachen Sequenzen und/oder verallgemeinerten Sequenzen erstellt. Dieser hoch performante Ansatz hat jedoch einen deutlichen Nachteil: in frühen Schritten eingebaute Fehler können im Verlauf der Berechnung nicht mehr verbessert werden. So werden falsch eingefügte Gaps nicht mehr entfernt. Für Sequenzdatensätze mit keiner zu geringen Sequenzähnlichkeit und keiner zu starken Fragmentierung führt dieser Ansatz jedoch zu sehr guten Ergebnissen.

CLUSTALW erzeugt das Alignment in vier Schritten: (1) Berechnung der Distanzen für alle Sequenzpaare über die Sequenzähnlichkeit, (2) Erstellung eines Nachbarschafts-Baums aus den Distanzen über die Neighbour-Joining Methode, (3) Berechnung von Sequenzgewichtungen, (4) Berechnung des progressiven Alignments entsprechend dem Nachbarschafts-Baums.

Für die Berechnung der Distanzen wurde die in CLUSTALW implementierte Methode der dynamischen Programmierung zum Vergleich von Sequenzpaaren gewählt. Für die Berechnung der Distanzwerte für einfache Sequenzpaare wurde die Gonnet Distanzmatrize sowie ein GOP (Gap Opening Penalty) von 10 und ein GEP (Gap Extension Penalty) von 0.1 gewählt. Für den progressiven Aufbau des Multisequenz Alignments wurde die Gonnet Distanzmatrize sowie ein GOP von 10 und ein GEP von 0.2 gewählt.

5.7.3 Erstellung der LED90 Datenbank

Für die Bestimmung der konservierten Bereiche innerhalb der Superfamilien, als auch die Erzeugung der HMM Profile ist es zu vermeiden, dass einzelne Sequenzcluster mit hoher Sequenzidentität überrepräsentiert sind und sich somit eine Überbewertung dieser Sequenzen ergibt. Deshalb wurden diese Untersuchungen mit einer nicht redundanten Version der LED durchgeführt, deren Sequenzen eine maximale Sequenzidentität von 90% aufwiesen (LED90). Diese Version wurde mit dem in CD-HIT implementierten Algorithmus erstellt.[156,157] Dieser sortiert zuerst sämtliche Sequenzen nach der Länge in absteigender Ordnung. Die längste Sequenz wird als repräsentative Sequenz des ersten Clusters gewählt. Dann wird jede verbleibende Sequenz mit den vorhandenen repräsentativen Sequenzen verglichen und wenn die Ähnlichkeit über einen gewählten Schwellenwert (90%) liegt, wird diese in dem zugehörigen Cluster abgelegt. Ansonsten wird diese als repräsentative Sequenz eines neuen Clusters bestimmt.

Der langsamste Schritt ist hierbei die Bestimmung der Sequenzähnlichkeit über ein paarweises Alignment der Sequenzen. Um dieses zu Beschleunigen ist in CD-HIT ein Filtermechanismus sowie eine Index Tabelle implementiert. Der Filter basiert auf der Tatsache, dass zwei Sequenzen die z.B. 85% Identität aufweisen in einem Fenster von 100 Aminosäuren mindestens 70 identische Dipeptide, 55 Tripeptide und 25 Pentapeptide besitzen. Deshalb müssen für Sequenzpaare, die diese Bedingung nicht erfüllen keine Alignments erstellt werden, was den Programmablauf beschleunigt.

5.7.4 Berechnung der für das Lösungsmittel zugängliche Proteinoberfläche

Die für das Lösungsmittel zugängliche Proteinoberfläche wurde mit dem in NACCESS 2.1.1 implementierten Algorithmus berechnet.[158] NACCESS nutzt die von Lee und Richards eingeführte Methode.[159] Das Lösungsmittel wird als ein kugelförmiges Molekül mit konstantem Radius r abstrahiert (r = 1,4 Å für Wasser). Beim Rollen dieser kugelförmigen Sonde über das Protein, beschreibt das Zentrum der Sonde die zugängliche Oberfläche. Neben diesen für jedes Atom berechneten absoluten Werten bestimmt NACCESS ebenso die aufsummierte, absolute sowie die relative Zugänglichkeit für jede Aminosäure. Die relativen Werte werden berechnet als %-Zugänglichkeit der Aminosäure x im Vergleich zur Zugänglichkeit dieser Aminosäure in einem Ala-x-Ala Tripeptid.[160] Aminosäuren mit einer relativen Zugänglichkeit kleiner 5% wurden als im Inneren des Proteins lokalisiert und nicht für das Lösungsmittel zugänglich definiert.[161]

5.7.5 Bestimmung der Sekundärstruktur von Proteinen

Zur Bestimmung der Sekundärstrukturelemente für Proteinstrukturen wurde der in DSSP implementierte Algorithmus verwendet.[94] Die Methode basiert vornehmlich auf der Bestimmung von H-Brücken Mustern zwischen Peptideinheiten. H-Brücken können in Proteinen über ein elektrostatisches Modell beschrieben werden, wobei eine energetisch günstige H-Brücke eine Bindungsenergie von etwa -3 kcal/mol besitzt. DSSP erlaubt jedoch eine relativ große Abweichung von diesem Idealwert und definiert H-Brücken über Energien die geringer als -0,5 kcal/mol sind. DSSP benutzt jedoch nicht direkt dieses Einparameter Model sondern nutzt eine geometrische Beschreibung über den Winkel θ zwischen den Atomen H-N···O und dem Abstand N···O, das mit dem Einparameter Modell korreliert. Eine ideale H-Brücke besitzt d = 2,9 Å, θ = 0 und E = -3 kcal/mol. Mit einer Höchstenergie von -0,5 kcal/mol erlaubt DSSP bei idealem Abstand eine Abweichung für θ bis 63° und bei idealem Winkel einen Abstand von d = 5,2 Å.

Minimale Helices der Länge n (n = 3, 4, 5) von Aminosäure i bis Aminosäure $i + n - 1$ werden definiert über die H-Brücke ($i - 1$, i + n – 1) und der H-Brücke (i, i + n). Längere Helices werden durch Überlagerung von minimalen Helices definiert. Für die Bestimmung von Strands wurde ein Brückenkonzept definiert:

parallele Brücken := [H-Brücke ($i - 1$, j) und H-Brücke (j, $i + 1$)] oder [H-Brücke ($j - 1$, i) und H-Brücke (i, $j + 1$)]

antiparallele Brücken := [H-Brücke (i, j) und H-Brücke (j, i)] oder [H-Brücke ($i - 1$, $j + 1$) und H-Brücke (j - 1, $i + 1$)]

5.7.6 Profil HMM-HMM Vergleich

Für die Erzeugung der HMM Profile sowie die Profil HMM-HMM Vergleiche wurden die in HHmake und HHsearch implementierten Algorithmen verwendet.[9] Hidden Markov Modelle (HMMs) stellen eine geschlossene Theorie zur Beschreibung von Profilen dar. HMMs umfassen eine Gruppe von wahrscheinlichkeitsbasierten Modellen die zur Beschreibung von linearen Sequenzen, bestehend aus Symbolen, geeignet sind. HMMs wurden in den späten 80ern in die Bioinformatik eingeführt[162] und ein paar Jahre später erstmals zur Beschreibung von Profilen eingesetzt.[24]

Ein HMM besteht aus einem Satz von Zuständen einschließlich einem Startzustand T_0 und einem Endzustand T_m. Für jedes Zustandspaar (T_i, T_j) existiert eine Wahrscheinlichkeit $P(i, j)$ um vom Zustand T_i in den Zustand T_j überzugehen. Die Wahrscheinlichkeiten sind Parameter,

die gewählt werden können. Die daraus ableitbare Wahrscheinlichkeitsverteilung muss jedoch folgende Randbedingung erfüllen:

$$0 \le P(i, j) \le 1 \text{ für alle } (i, j)$$

$$\sum_j P(i, j) = 1$$

Durch ein HMM können verschiedene Wege $\Pi = \pi_1 ... \pi_n$ gewählt werden, wobei $\pi_1 = T_0$ und $\pi_n = T_m$ ist. Unter der Annahme, dass die Übergangswahrscheinlichkeiten der Zustände voneinander unabhängig sind ergibt sich die Wahrscheinlichkeit für einen gewählten Weg Π nach

$$P(\Pi) = \prod_{i=2,...,n} P(\pi_{i-1}, \pi_i).$$

Jeder Zustand kann ein Symbol emittieren und besitzt seine eigene Wahrscheinlichkeitsverteilung für das zu Grunde liegende Symbolalphabet (z.B. Aminosäuren). $P(a \mid j)$ beschreibt die Wahrscheinlichkeit, dass Zustand T_j das Symbol a erzeugt. Diese Wahrscheinlichkeiten sind wählbare Parameter müssen aber folgende Randbedingung erfüllen:

$$0 \le P(a \mid j) \le 1 \text{ für alle } a \text{ und } j$$

$$\sum_a P(a \mid j) = 1 \text{ für alle j.}$$

Ein HMM erzeugt somit eine Sequenz, indem es einem Weg Π folgend für jeden Zustand ein Symbol emittiert und die Verknüpfung dieser emittierten Symbole zur Sequenz führt. Die Wahrscheinlichkeit, dass einem spezifischen Weg Π folgend das HMM die Sequenz $q = q_1 ... q_n$ erzeugt ist somit

$$P(q, \Pi) = P(\Pi) P(q \mid \Pi) = \prod_{i=2,...,n} P(\pi_{i-1}, \pi_i) \prod_i P(q_i \mid \pi_i).$$

Die Profil HMMs für die homologen Familien der LED wurden mit dem in HHmake implementierten Algorithmus generiert.[9] Die hierfür nötigen Multisequenz Alignments wurden mit CLUSTALW für die Sequenzen aus der nr90 erstellt. Diese Alignments wurden als Training Set verwendet, um die Übergangs- und Emissionswahrscheinlichkeiten zu parametrisieren. Die so definierten Profil HMMs ermöglichen den Vergleich des HMMs mit einer Sequenz und ermöglichen zwischen Sequenzen, die Mitglied einer Familie sind und denen, die es nicht sind zu unterscheiden. Für den Vergleich der HMMs mit einer Sequenz müssen pro Spalte des Alignments drei Zustände definiert werden: ein *match* Zustand (M), ein *insertion* Zustand (I) und ein *delete* Zustand (D). Der Zustand M steht für die Übereinstimmung einer Aminosäure zwischen Sequenz und HMM, der D Zustand übergeht eine Spalte im HMM und emittiert ein Leerzeichen, der I Zustand erlaubt das Einfügen einer

oder mehrer Aminosäuren (das Einfügen mehrerer Aminosäuren ist möglich, da eine Übergangswahrscheinlichkeit des I Zustands in sich selbst besteht).

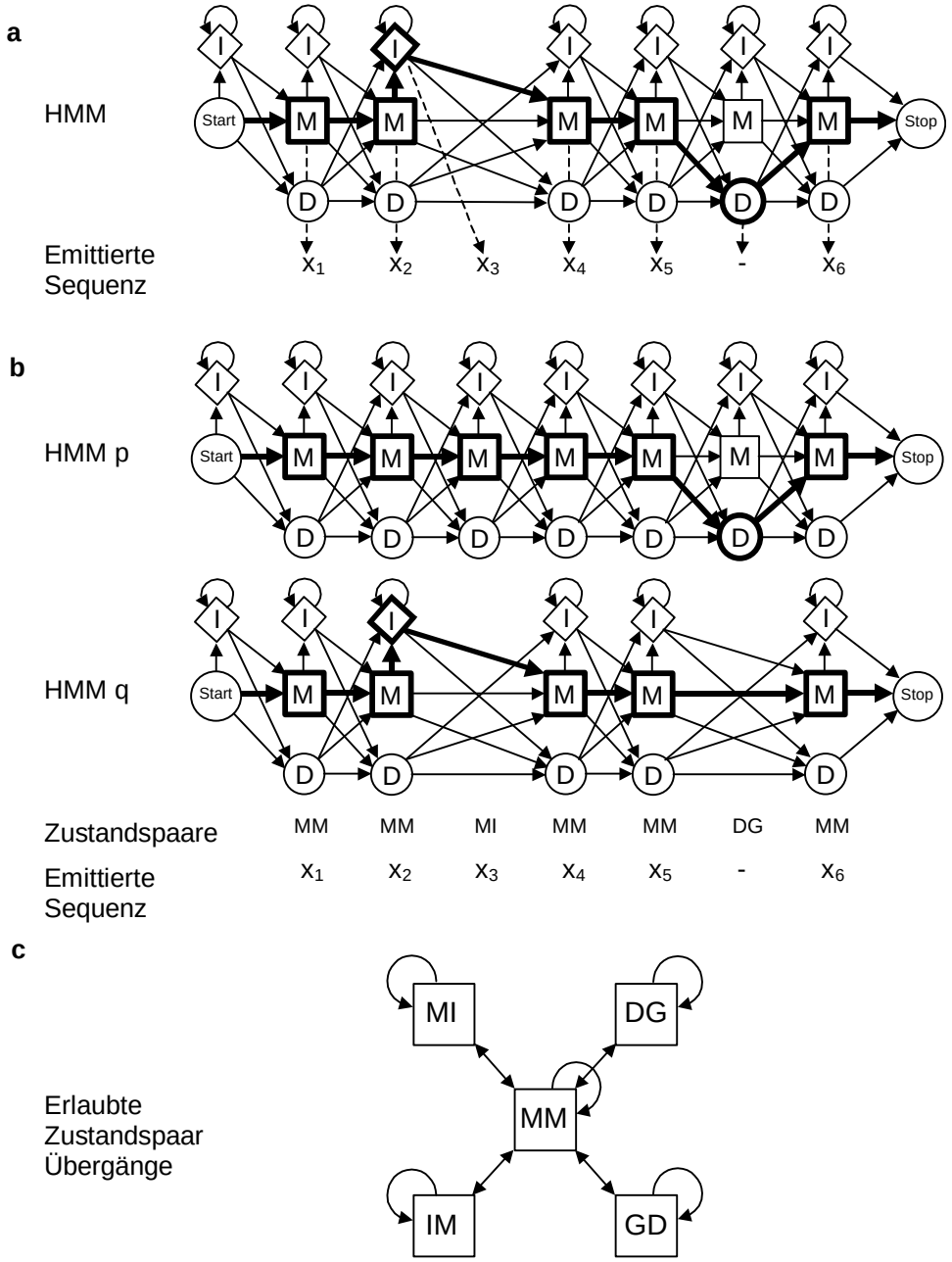


Abbildung 42: (a) Alignment eines Profil HMM mit einer Sequenz. (b) Alignment zweier HMMs. Der Pfad durch die zwei HMMs (fett) entspricht der einer Sequenz, die durch beide HMMs emittiert wird. (c) Erlaubte Übergänge zwischen Profil HMM-HMM Zustandspaaren.[9]

Für den Profil HMM-HMM Vergleich wurde der in HHsearch implementierte Algorithmus verwendet.[9] Dieser erstellt ein Alignment für zwei HMMs durch Maximierung der Coemissionswahrscheinlichkeit.[163] Bei einem Alignment zweier HMMs ist zu beachten, dass M und I Zustände eines HMMs nur mit einem M oder I Zustand des zweiten HMMs gepaart

werden dürfen. Ebenso darf ein D Zustand nur mit einem D Zustand oder einem *Gap* (G) gepaart werden. Diese Zustandspaare werden als MM, MI, IM, II, DD, DG und GD bezeichnet. In Abbildung 42b ist ein Beispiel für ein Profil HMM-HMM Alignment dargestellt. In der dritten Spalte emittiert HMM p eine Aminosäure aus einem M Zustand und HMM q emittiert eine Aminosäure aus einem I Zustand. Das Zustandspaar für diese Spalte im Alignment ist somit MI. In der sechsten Spalte wird aus HMM p nichts emittiert, da es einen D Zustand durchläuft. HMM q emittiert ebenfalls nichts, da sich an dieser Stelle im Alignment ein *Gap* befindet. Dieses Zustandspaar entspricht DG. Es sind nur Übergänge von Zustandspaaren in sich selbst und aus den Zustandspaaren MI, IM, DG oder GD in das Zustandspaar MM erlaubt (Abbildung 42c).

Sowohl für Profil HMM-Sequenzvergleiche als auch für Profil HMM-HMM Vergleiche muss der Pfad durch das HMM gefunden werden, der die höchste Wahrscheinlichkeit besitzt. In HHsearch ist hierzu der Viterbi Algorithmus implementiert, der auf dynamischer Programmierung basiert.

5.7.7 Aminosäure Konservierung

Für die Berechnung der Aminosäure Konservierung C innerhalb der LED Superfamilien wurde der in AL2CO implementierte Algorithmus verwendet.[164] Es wurde die *sum of pairs* Methode mit ungewichteten Aminosäurehäufigkeiten gewählt. Die Konservierung C wurde nicht normalisiert. Diese Methode verwendet Informationen über Mutationen aus Substitutionsmatrizen um stereochemische Variationen innerhalb einer Spalte eines Alignments zu bewerten. Für die hier beschriebenen Untersuchungen wurde die BLOSUM62 Matrize gewählt. Die Alignments der Superfamilien wurden mit CLUSTALW für die Sequenzen aus der LED90 erstellt. Hierbei wurden nur Superfamilien berücksichtigt, für die mindestens 20 Sequenzen in LED90 vorhanden waren, sowie eine Proteinstruktur zur Verfügung stand.

Die Berechnung der Konservierung erfolgte in zwei Schritten. Zuerst wurden ungewichtete Aminosäurehäufigkeiten bestimmt: $f_a(i) = n_a(i) / n(i)$, wobei $n_a(i)$ der Anzahl der Sequenzen entspricht für die Position i mit Aminosäure a besetzt ist und $n(i)$ der Anzahl der Sequenzen im Alignment entspricht für die Position i vorhanden ist (ein Gap an dieser Position wird nicht berücksichtigt):

$$n(i) = \sum_{a=1}^{20} n_a(i) .$$

Im zweiten Schritt wurde die Konservierung C aus den Aminosäurehäufigkeiten mit der *sum of pairs* Methode berechnet:

$$C(i) = \sum_{a=1}^{20} \sum_{b=1}^{20} f_a(i) f_b(i) S_{ab} ,$$

wobei S_{ab} einem Wert der BLOSUM62 Matrize entspricht. Die Konservierung C ist größer für Positionen, die mit ähnlichen Aminosäuren besetzt sind. Da die diagonalen Werte der Substitutionsmatrizen nicht identisch sind, unterscheidet sich die Konservierung C für invariante Positionen in Abhängigkeit von der Art der Aminosäure. Im Fall der BLOSUM62 Matrize wird der höchste Wert von 11 für Tryptophan erreicht.

5.8 Die Lipase Engineering Database (LED)

Die für die Erstellung der LED verwendeten Suchsequenzen wurden der in SCOP[63] beschriebenen Superfamilie „a/b hydrolase fold (53473)“ entnommen und um weitere Sequenzen ergänzt, die durch Stichwortsuchen in GenBank[92] ermittelt wurden. Insgesamt ergaben sich somit 61 Suchsequenzen (Anhang B).

Mit diesen Sequenzen wurden Homologiesuchen gegen den nicht redundanten Datensatz der GenBank durchgeführt. Hierbei wurde als Schwellenwert ein E-Wert von 10^{-10} gewählt. Für die Klassifikation durch den automatisierten Vergleich gegen die lokale BLAST Datenbank wurde als Schwellenwert für die Zugehörigkeit zu einer homologen Familie ein E-Wert von 10^{-40} und für die Zugehörigkeit zu einer Superfamilie ein E-Wert von 10^{-10} gewählt. Eine Proteinsequenz wurde einem existierenden Proteineintrag zugewiesen, falls die Sequenzidentität über 98% lag.

Für die Überlagerung der Proteinstrukturen wurden die C_α-Atome folgender im α/β-Hydrolasefold konservierten Bereiche gewählt: der Aminosäuren um das katalytisch aktive Nucleophil Nu (Nu_{-6}-Nu-Nu_{+2}), des katalytisch aktiven Histidins und dessen direkten Nachbarn, der katalytisch aktiven Säure sowie vier weiterer gut konservierter Aminosäuren des zentralen β-Strands β6.

5.9 Bestimmung der konservierten Positionen SKP und ZKP

Für die Bestimmung der konservierten Positionen des α/β-Hydrolase Folds wurden nur Superfamilien herangezogen, die eine bekannte Proteinstruktur und mindestens 20 Proteinsequenzen mit einer Sequenzähnlichkeit kleiner 90% enthielten. Für diese Familien

wurde mit CLUSTALW ein Multisequenz Alignment erstellt und mit dem in AL2CO implementierten Algorithmus die Konservierung C dieser Familien bestimmt. Um die Konservierung zwischen diesen Superfamilien vergleichen zu können, wurde ein Strukturalignment der Referenzstrukturen mit dem Swiss PDB Viewer erstellt. Für die Überlagerung wurden 29 C_α-Atome der zentralen β-Strands β3-β7 und der Crossover Helices αA-αF verwendet. Nach der manuellen Überarbeitung des Strukturalignments wurde für den Satz der strukturell überlagerten Positionen die Konservierung dieser Aminosäuren innerhalb der Superfamilien miteinander verglichen. Hierbei wurden zuerst alle Positionen die eine Konservierung C größer dem Durchschnittswert der entsprechenden Superfamilien-konservierung aufwiesen bestimmt und als zwingend konservierte Positionen (ZKP) definiert. Die Aminosäuren, die eine Konservierung C größer Null aufwiesen wurden als streng konservierte Positionen (SKP) definiert (Anhang B). Somit stellten die ZKPs eine Teilmenge der SKPs dar. Der Schwellenwert für die Konservierung der SKPs wurde mit Null relativ niedrig angesetzt, da durch das Kriterium, dass alle untersuchten Superfamilien dem Schwellenwert genügen mussten, die Anzahl der zu erwartenden Positionen bereits stark eingeschränkt wurde.

5.10 Form der Bindungstasche

Um die Form der Bindungstaschen der α/β-Hydrolasen zu vergleichen, wurden die Connolly Oberflächen[165] mit einem Probenradius von 1,5 Å für die geöffneten Proteinstrukturen berechnet, wobei vorhandene Inhibitoren von der Berechnung ausgeschlossen wurden. Zur Visualisierung der Oberflächen wurde InsightII (Molecular Simulations, San Diego, CA) verwendet. Für jede Porteinstruktur wurden drei senkrecht zueinander stehende Ansichten erzeugt, die Querschnitte mit 2 Å Breite darstellten, die das Phosphoratom (im Falle der *Candida rugosa* Lipase das Schwefelatom) des gebundenen Inhibitors enthielten. Für alle untersuchten Strukturen wurden die selben Ebenen verwendet, die aus der Struktur der *Candida antarctica* Lipase B abgeleitet wurden.

Für die Seitenansicht wurde eine Ebene durch drei Inhibitoratome gelegt (Phosphoratom, sowie zwei benachbarte Atome, Kohlenstoff und Estersauerstoff) (Abbildung 30a). Diese Ansicht visualisiert die Inhibitoratome nahe des katalytisch aktiven Nucleophils. Die Frontansicht ist eine zur Seitenansicht senkrecht stehende Ebene, die durch das Phosphoratom, das Sauerstoffatom der Seitenkette des katalytisch aktiven Nucleophils und dem Oxyanionanalogons gelegt wurde. Diese Ansicht verläuft entlang der Achse des

Inhibitors oder des Substrats vom Alkoholrest zum Acylrest. Die Aufsicht steht senkrecht zur Front- und Seitenansicht und visualisiert die Sicht in die Bindungstasche hinein. Schematische Ansichten wurden manuell erstellt, wobei die Connolly Oberflächen durch gerade Linien und Kreissegmente repräsentiert wurden.

5.11 Modelle der Fettsäurekomplexe

Für die Bestimmung der Eigenschaften der Lipase Bindungstaschen wurden Proteinstrukturen mit dem Programm GRID[166] untersucht. Unter Verwendung der METHYL und DRY Sonden konnten enge und hydrophobe Bereiche identifiziert werden. Falls in der Proteinstruktur ein Inhibitormolekül vorhanden war, wurde es vor den Berechnungen entfernt. Die Interaktionspotentiale, die durch GRID berechnet werden, sind eine Kombination aus sterischer Abstoßung, elektrostatischen Wechselwirkungen und Wasserstoffbrücken. Dadurch erlaubt GRID eine quantitative Analyse der Eigenschaften zur Bindung von chemischen Verbindungen an der Proteinoberfläche. Isopotentialflächen für –2,2 kcal/mol (METHYL Sonde) und –0,22 kcal/mol (DRY Sonde) wurden mit Insight II visualisiert. Schematische Darstellungen der potentiellen Bindungstasche der Fettsäurekette wurden manuell erstellt. Hierfür wurden die vorhergesagten Bindungsbereiche durch gerade Linien manuell repräsentiert.

Modelle für gesättigte Methylfettsäureester wurden in ihrem tetrahedralen Übergangszustand mit SYBYL (Tripos, St. Louis, MO) erstellt. Durch Überlagerung des tetrahedralen Kohlenstoffatoms der polarisierten Carbonylgruppe mit den Phosphor- oder Schwefelatomen der Inhibitoren der Proteinstrukturen, wurden die Modelle in der Bindungstasche platziert. Die Geometrie der Fettsäurekette wurde manuell an die Konformation des Inhibitors angepasst. Fettsäureketten die länger als der Inhibitor waren, wurden zuerst entlang des Inhibitors gelegt und dann entlang der Bindungsbereiche, die von der METHYL Sonde vorhergesagt wurden und schließlich entlang den Bindungsbereichen der DRY Sonde. Wenn möglich wurden die Torsionswinkel der trans-Konformation gewählt. Der Aufbau der Fettsäurekette wurde abgebrochen, sobald das Ende des von GRID vorhergesagten hydrophoben Bindungsbereichs oder das Ende der Bindungstasche erreicht wurde.

5.12 Familienspezifische Primer

Für das Design der familienspezifischen Primer wurden aus den Proteineinträgen der Familie abH1 zwei Sequenzdatensätze S_1 und S_2 ermittelt wobei nur Proteine mit Hydrolaseaktivität ausgewählt wurden. S_1 enthielt alle in der Familie abH1 vorhandenen Proteinsequenzen mit Hydrolaseaktivität, die ursprünglich aus Swiss-Prot extrahiert wurden (33 Sequenzen). S_2 setzte sich aus 12 Referenzsequenzen zusammen, die den Sequenzraum der Familie abH1 wiedergaben. Diese Sequenzen tragen folgende Swiss-Prot Bezeichner: EST1_CULPI, ESTE_MYZPE, PNBA_BACSU, PCD_ATROX, CRYS_DICDI, EST2_CAEEL, EST1_CAEBR, ESTJ_HELVI, BAL_RAT, EST1_RAT, ESTB_DROPS und ACES_TORCA.

Konservierte Sequenzblöcke wurden für beide Sequenzdatensätze mit den in BLOCK MAKER[20] implementierten Methoden MOTIF[167] und Gibbs Sampler[168,169] bestimmt. Identifizierte Sequenzblöcke wurden als signifikant gewertet und für das Primerdesign berücksichtigt, wenn diese sowohl von der MOTIF als auch der Gibbs Sampler Methode identifiziert wurden.

MOTIF identifizierte für S_1 4 Sequenzblöcke ($1M_1$-$1M_4$) (Anhang C), Gibbs Sampler 6 Blöcke ($1G_1$-$1G_6$) (Anhang D). MOTIF und Gibbs Sampler fanden gemeinsame Sequenzblöcke für den Bereich zwischen β-Strand β1 und β2 ($1M_2$, $1G_1$), den β-Strand β2 einschließlich den 4 N-terminal liegenden Aminosäuren ($1M_3$, $1G_2$) und den Bereich von β5 bis β6 der das GXSXG Motiv einschließt ($1M_4$, $1G_5$).

MOTIF identifizierte für S_2 5 Sequenzblöcke ($2M_1$-$2M_5$) (Anhang E), Gibss Sampler 6 Blöcke ($2G_1$-$2G_6$) (Anhang F). MOTIF und Gibbs Sampler fanden gemeinsame Sequenzblöcke für den Bereich zwischen β-Strand β1 und β2 ($2M_2$, $2G_1$), den β-Strand β2 einschließlich den 4 N-terminal liegenden Aminosäuren ($2M_3$, $2G_2$), das *oxyanion hole* ($2M_3$, $2G_3$), den Bereich um β-Strand β4 (rM_4, rG_4) sowie den Bereich von αA bis αC, der das GXSXG Motiv einschließt (rM_5, rG_5+rG_6).

Die so identifizierten Blöcke wurden mit CODEHOP[170] für das Design von Hybrid Primer prozessiert. Die Standardeinstellungen des CODEHOP Webinterface für die Erkennung des *Core*s (degeneracy = 128, strictness = 0), die Konstruktion der *Clamp* Region (temperature = 60°C, poly-nuc = 5), die Primer Konzentration (50 nM) und den genetischen Code (standard) wurden übernommen. Um den unterschiedlichen GC Gehalten der verschiedenen Genome gerecht zu werde, wurde der *Clamp* Bereich sowohl unter Verwendung der Codon Usage von *Homo sapiens* sowie *Bacillus subtilis* generiert.

Aus den für den Datensatz S_1 abgeleiteten Blöcken konnte CODEHOP nur für $1M_3$ und $1G_2$, die den Bereich um den β-Strand β2 einschließen Hybrid Primer mit einer Degenerierung nicht größer als 128 konstruieren. Es wurden für jede Codon Usage drei reverse Primer vorgeschlagen deren *Core* Region aus den 4 N-terminal vorgelagerten Aminosäuren des β-Strands β2 mit Degenerierungen von 16 bzw. 128 gebildet wurde.

Aus den für den Datensatz S_2 abgeleiteten Blöcken konnte CODEHOP nur für $2M_5$ und $2G_5$, die den Bereich von αA bis αC umfassen, Hybrid Primer mit einer Degenerierung nicht größer als 128 konstruieren. Es wurde für jede Codon Usage jeweils ein vorwärts und ein reverser Primer vorgeschlagen deren *Core* Region mit einer Degenerierung von 64 aus 4 Aminosäuren, die dem β-Strand β5 N-terminal vorgelagerten sind, bestand.

Die vorwärts Primer wurde aus den CODEHOP Ergebnissen für den Datensatz S_1 abgeleitet. Jedoch war der N-terminale Bereich des identifizierten *Core* Bereichs nur schlecht erhalten. Somit konnten nur reverse Primer für die Blöcke $1M_3$ und $1G_2$ generiert werden. Deshalb wurde unter Beibehaltung des *Core*s die *Clamp* Region manuell designt. Hierzu wurde DNA Sequenzen mit verschiedenen GC Gehalt bestehend aus Restriktionsstellen konstruiert. Diese Vorgehensweise ist möglich, da die eigentliche Funktion des *Clamp* Bereichs in der Erhöhung der Annealing Temperatur besteht und degenerierte Primer mit *Clamp* Bereichen die nur eine geringen Übereinstimmung aufweisen bereits erfolgreich eingesetzt wurden.[170] Dies führte zu den Primern v1 und v2.

Die reversen Primer wurden aus den CODEHOP Ergebnissen für den Datensatz S_2 abgeleitet. Die für die Sequenzblöcke $2M_5$ und $2G_5$ unter Verwendung sowohl der *Bacillus subtilis* als auch *Homo sapiens* Codon Usage erhalten reversen Primer wurden direkt übernommen. Dies führte zu den Primern rBS und rHS.

Des Weiteren wurde zur Evaluierung die den vorwärts und reversen Primern entsprechenden Bereiche in der pNBA aus *Bacillus subtilis* als Primer eingesetzt (v-pNBA, r-pNBA)

Tabelle 10: Primer für familienspezifische Amplifikation

Primer	Primersequenz
v1	5'-CTGCAGGAATTCCGCCAGTCCGARGAYTGYYT-3'
v2	5'-CTGCAGGAATTCGCGGCCGCGARGAYTGYYT-3'
EDCL3	5'-CTGCAGGAATTCCGCCAGTCCGAGGATTGCTTG-3'
rBS	5'-GGCCAAAAATTGTAATATTATTCGGATYNCCNCCRAA-3'
rHS	5'-GATGGTGATGTTGTTGGGGTYNCCNCCRAA-3'
v-pNBA	5'-GAGGATTGCTTGTATGTCAATGTATTTGCGCCTGACAC-3'
r-pNBA	5'-TCCAAATACTGTTACGTTATCGGGATCACCGCCAAACG-3'

5.13 Geräte und Verbrauchsmittel der experimentellen Arbeiten

Soweit nicht anders vermerkt, sind die jeweiligen Hersteller bzw. Vertriebsniederlassungen in Deutschland ansässig. Neben den hier aufgeführten Arbeitsmitteln und Geräten wurden gängige Laborgegenstände (Glasgeräte, Pipetten, Magnetrührer usw.) verwendet.

5.13.1 Biotransformation

Küvetten	1 ml-Plastikküvetten (Greiner, Frickenhausen), 1 ml-Glasküvetten (Hellma Optik, Jena)
pH-Meter	Digital pH-Meter 525 (WTW, Weilheim), pH-Meter 620 (Metrohm, Herisau, Schweiz)
Thermomixer	Thermomixer comfort (Eppendorf, Hamburg), Schutron Thermoshaker (Neolab, Heidelberg)
Waagen	MC1 Research RC210D (Sartorius, Göttingen), HF2000G (A&D Engineering, Milpitas, USA)

5.13.2 Mikrobiologie, Molekulargenetik und Proteinaufreinigung

Agarose-Gelelektrophorese	DNA Sub Cell, Mini DNA Sub Cell (BioRad, München), Video Copy Processor P66E (Mitsubishi, Cambridge, Großbritannien), BWM9X Monitor (Javelin Electronics, Schaumburg, USA), UV-Leuchttisch (MWG-Biotech, Ebersberg)
Autoklav	PACS 2000 (Getinge AB, Schweden)

Brutschränke	WTE (Binder, Tuttlingen), UM500 (Memmert, Schwabach)
Dialyse	Spectra/Por-Dialyseschläuche (Spectrum Medical Industries), PD10-Chromatographiesäulen (Amersham Pharmacia Biotech, Uppsala, Schweden)
DNA-Sequenzierung	ABI Prism 377 DNA Sequencer mit Auswertungssoftware ABI Prism 377 Version 3.2 (PE Applied Biosystems, Weiterstadt)
Inkubatoren	Mutitron HT Schüttler (Infors, Bottmingen, Schweiz)
PCR-Geräte	Master Cycler Gradient (Eppendorf, Hamburg)
Photometer	Für Küvetten (1 cm): UV/VIS-Spektrometer Ultrospec3000 und BioChrom4060 (Amersham Pharmacia Biotech, Uppsala, Schweden) Zur Bestimmung der DNA-Konzentration: BioPhotometer (Eppendorf, Hamburg) inkl. UVette-Küvetten 220-1600 nm
SDS-PAGE	Minigel-Twin G42 (Biometra, Göttingen), Model 583 Gel Dryer (Biorad Laboratories, München), Wipptisch ROCKY (Labor-technik Fröbel, Lindau)
SpeedVac	Concentrator 5301 (Eppendorf, Hamburg)
Ultraschallgerät	Sonifier 250 (Branson, Danbury, USA)

Vortexer	VF2 (IKA Labortechnik, Staufen), IKA Minishaker MS1 (IKA Labortechnik, Staufen)
Zentrifugen	Centrifuge 5416, 5417C, 5417R und 5810R mit Einsätzen für Kunststoffgefäße (0,5 ml; 1,5 ml; 2 ml; 15 ml; 50 ml) und Mikrotiterplatten (Eppendorf, Hamburg), Sorvall RC-5B und RC-5C Refrigerated Superspeed Centrifuge (DuPont Instruments, Leipzig) mit Rotoren SLA3000, SA600 und SS34

5.13.3 Chemikalien und Enzyme

BioRad Laboratories (München)	SDS-PAGE Standard (Broad Range), 1kB-DNA-Längen-standard
DIFCO-Laboratories (Detroit, USA)	Trypton, Hefeextrakt
Eppendorf (Hamburg)	DNA-*Taq*-Polymerase (5 U/µl) incl. 10x *Taq* -Puffer und $MgCl_2$-Lösung
Fluka Chemie (Buchs, Schweiz)	Aceton, Agar, Coomassie Brilliant Blue R-250, Dextran Blau, Dichlormethan, Dioxan, DMSO, Ethanol, Ethidiumbromid-Lösung (1% in Wasser), Glycerol, H_2SO_4 konz., HCl konz., Hefe-Extrakt, KCl, Lysozym (aus Hühner-eiweiss), NaCl, NaOH, NaOAc, Na_2SO_4 Natriumpyrophosphat, SDS, Tetrahydrofuran, Triethylamin
Gibco BRL (Eggenstein)	dNTPs, Agarose

Life Technologies GmbH (Karlsruhe)	10xTBE-Puffer
MBI Fermentas (St. Leon-Roth)	Restriktionsendonukleasen, T4 DNA-Ligase inkl. 10xPuffer
Perkin Elmer (Weiterstadt)	BigDye Terminator Ready Reaction Kit für DNA-Sequenzierung
Pierce (St. Augustin)	BCA-Kit zur Bestimmung der Protein-konzentration
Qiagen GmbH (Hilden)	QIAprepspin Miniprep Kit, QIAquick Gel Extraction Kit, Qiagen Plasmid Midi Kit
Roth GmbH (Karlsruhe)	Ampicillin, Rotiphorese NF-10xTBE-Puffer, Rotiphorese NF-Harnstoff, Rotiphorese NF-Acrylamid / Bisacrylamid-Lösung 40%, Acrylamid-Lösung für SDS-PAGE, Phenol / Chloroform - Lösung, Trypton, Hefeextrakt
Riedel-de-Haen (Seelze)	Tris-(hydroxymethyl)-aminomethan
Serva Feinbiochemika GmbH (Heidelberg)	Agarose
Sigma-ARK GmbH (Darmstadt)	DNA-Oligonukleotide

5.14 Verwendete Mikroorganismen und Plasmide

Folgende Stämme wurden vom DSMZ erhalten und zur Isolierung von genomischer DNA verwendet, die als Matrize für die Untersuchung der familienspezifischen Primer diente: *Aspergillus nidulans*, *Bacillus subtilis* Stamm 168, *Bacillus firmus*, *Bacillus megaterium*, *Candida rugosa*, *Streptomyces coelicolor* und *Streptomyces lividans*. Der *E. coli*-Stamm DH5α (*supE*44, *lacU*169 (80*lacZ* M15) *hsdR*17 *recA*1 *endA*1 *gyrA*96 *thi*-1 *relA*1) wurde bei Clontech (Heidelberg) erworben und für Klonierungsarbeiten und Expression verwendet. Als

Vektoren dienten pCR®2.1-TOPO® (Invitrogen) zur Klonierung der familienspezifischen Primer Produkte, sowie pCYTEXP1[171] und pT-ompA-Δ70HpHis[172] mit dem Hitze induzierbaren λ Promotor zur Expression der *Burkholderia cepacia* Lipase (BCL) Varianten bzw. des BCL Chaperon. Die im Rahmen dieser Arbeit zur heterologen Expression von BCL-Enzymen verwendeten Plasmide sind in Tabelle 11 zusammengefasst.

Tabelle 11: Verwendete Plasmide

Plasmid	Selektionsmarker	Eingeführte Mutationen
pT-BCL	Ampicillin-Resistenz	Wildtyp
pT-BCL-L17T		L17T
pT-BCL-L17T-L167D		L17T L167D
pT-BCL-L17T-L167E		L17T L167E
pT-BCL-L17T-L167N		L17T L167N
pT-BCL-L17T-L167Q		L17T L167Q

5.15 Synthetische Oligonucleotide

5.15.1 Primer für familienspezifische Amplifikation

v1: 5’-CTGCAGGAATTCCGCCAGTCCGARGAYTGYYT-3’
v2: 5’-CTGCAGGAATTCGCGGCCGCGARGAYTGYYT-3’
EDCL3: 5'-CTGCAGGAATTCCGCCAGTCCGAGGATTGCTTG-3’
rBS: 5'-GGCCAAAAATTGTAATATTATTCGGATYNCCNCCRAA-3’
rHS: 5'-GATGGTGATGTTGTTGGGGTYNCCNCCRAA-3’
v-pNBA: 5'-GAGGATTGCTTGTATGTCAATGTATTTGCGCCTGACAC-3’
r-pNBA: 5'-TCCAAATACTGTTACGTTATCGGGATCACCGCCAAACG-3’

5.15.2 Primer für die QuikChange® PCR

v-L17T 5’-CACGGG**ACG**ACGGGTACCGACAAATACGCA-3’

v-L167D	5’-GCGCTTAAGACG**GAT**ACGACCGCGCAG-3’
v-L167	5’-GCGCTTAAGACG**GAA**ACGACCGCGCAG-3’
v-167N	5’-GCGCTTAAGACG**AAC**ACGACCGCGCAG-3’
v-L167Q	5’-GCGCTTAAGACG**CAG**ACGACCGCGCAG-3’

5.16 Lösungen, Medien und Puffer

Soweit nicht anders vermerkt, können alle Medien, Puffer und Lösungen bei Raumtemperatur gelagert werden.

5.16.1 Kulturmedien

Soweit nicht anders angegeben, wurden sämtliche Medien 30 min bei 121°C autoklaviert. Zur Herstellung fester Nährböden wurden die Medien vor dem Autoklavieren mit 15 g/l Agar versetzt. Das zur Selektion benötigte Ampicillin wurde nach Abkühlen der Nährlösungen auf ca. 45°C zugegeben, bei Flüssigkulturen erst kurz vor der eigentlichen Anzucht (Endkonzentration 100 µg/ml). Die Glucose wurde als 20%-ige Lösung separat sterilfiltriert und direkt vor der Inokulation zugegeben.

Luria-Bertani (LB) Medium:	Trypton	10 g
	Hefeextrakt	5 g
	NaCl	10 g
	dH_2O ad 1 l	
	pH-Wert	7,5

5.16.2 Puffer und Lösungen für Mini-Plasmid-Präparation (Schnelltest)

Resuspendierungspuffer:	Tris-HCl; pH 7,5	100 mM
(Lsg.I)	EDTA	10 mM
	RNAse I	400 µg/ml
Lysispuffer:	NaOH	1 M
(Lsg.II)	SDS	5,3%(m/v)
Neutralisationspuffer:	$MgCl_2$	3 M

(Lsg.III)	NaOAc	5 M

Lsg. I wurde autoklaviert, die RNAse erst direkt vor der Benutzung zugegeben.

5.16.3 Puffer und Lösungen für Agarose-Gelelektrophorese

TAE-Puffer (50x):	Tris	242 g
	Eisessig (99,8%)	57 ml
	0,5 M EDTA; pH 8,0	100 ml
	dH_2O	ad 1 l
6x-DNA-Probenpuffer:	Glycerin	30%(m/v)
	Bromphenolblau	0,2%(m/v)
	EDTA; pH 7,5	25 mM
Agarose-Gel (1%):	Agarose	4 g
(Lagerung bei 80°C)	TAE-Puffer	400 ml

5.16.4 Puffer und Lösungen für DNA-Sequenzierung

10%-APS-Lösung:	Ammoniumpersulfat	1 g
(Lagerung bei –20°C)	ddH_2O	10 ml
5,25%-Page-Plus-Gel:	Harnstoff	18 g
	TBE-Puffer (10x)	5 ml
	40%-Page-Plus-Lösung	6,6 ml
	ddH_2O	ad 50 ml
	TEMED	25 µl
	10%-APS-Lösung	250 µl
TBE-Puffer (10x):	Tris	108 g
	Borsäure	55 g

	EDTA	7,4 g

5.16.5 Puffer und Lösungen für SDS-Gelelektrophorese

Lower Tris (4x):	Tris	36,4 g
(Lagerung bei 6°C)	SDS	0,8 g
	dH_2O ad 200 ml	
	pH-Wert	8,8
Upper Tris (4x):	Tris	12,11 g
(Lagerung bei 6°C)	SDS	0,8 g
	dH_2O	ad 200 ml
	pH-Wert	6,8
10%-APS-Lösung:	s.o.	
SDS-Probenpuffer:	Glycerin	10 ml
	2-Mercaptoethanol	5 ml
	SDS	3 g
	0,05%-Bromphenolblau-Lsg.	2,5 ml
	Upper Tris (4x)	12,5 ml
	dH_2O	ad 50 ml
Elektrodenpuffer:	Tris	3 g
	Glycin	14,4 g
	SDS	1 g
	dH_2O	ad 1l
	pH-Wert	8,4
Coomassie-Färbelösung:	Coomassie-Brilliant-Blue	1g
	Essigsäure	100 ml
	Methanol	300 ml
	dH_2O	600 ml

Entfärbelösung:	Essigsäure	100 ml
	Methanol	300 ml
	dH_2O	600 ml
Proteinstandard: (Lagerung bei –20°C)	Rekombinante Proteine mit den Größen 10, 15, 25, 37, 50, 75, 100, 150 und 250 kDa	

5.16.6 Lösungen für die Transformation in *E. coli*

TSS-Medium:	LB-Medium	
(Lagerung bei 6°C)	PEG6000	10% (m/v)
	DMSO	5% (v/v)
	$MgCl_2$	50 mM

5.16.7 Lösungen für die QuikChange® PCR

Primer-Lösungen: (Lagerung bei –20°C)	Oligonukleotid	1 µM
dNTP-Mix:	dNTPs, 100 mM	je 2,5 µl
(Lagerung bei –20°C)	dH_2O	90 µl
10xPCR-Reaktionspuffer:	Tris	100 mM
(Lagerung bei –20°C)	$MgCl_2$	15 mM
	pH-Wert	8,3

5.16.8 Puffer und Lösungen für weitere Anwendungen

TE-Puffer:	Tris/HCl	10 mM
	EDTA	1 mM
	pH-Wert	7,5 – 8,5
Aufschlusspuffer I:	Tris	30 mM

	EDTA	50 mM
	NaCl	50 mM
	Lysozym	1 mg/ml
	pH-Wert	8,0
Aufschlusspuffer II:	KP_i	50 mM
	$MgCl_2$	10 mM
	Lysozym	1 mg/ml
	DNAse I	0,1 µg/ml
	pH-Wert	8,0

5.17 Mikrobiologische und molekulargenetische Methoden

Methoden die zur Validierung des familienspezifischen *Screenings* mit CODEHOP Hybridprimern notwendig waren, wurden von Jutta Schmitt angewandt.

5.17.1 Stammhaltung und Kultivierung von *Escherichia coli*

Zur Kultivierung von rekombinanten DH5α wurden 2-5 ml LB-Medium mit Ampicillin versetzt und mit einer Bakterienkolonie von einer LB-Amp-Agarplatte mittels sterilem Zahnstocher angeimpft. Die *E. coli*-Kultur wurde über Nacht bei 37°C unter Schütteln inkubiert (200 Upm). Diese Art der Kultivierung wird im Folgenden als *ÜN-Kultur* bezeichnet. Für Mini-Plasmid-Präparationen sind 2 ml ÜN-Kultur ausreichend, für Midi-Plasmid-Präparationen wurden 50 ml LB-Amp-Medium in sterilen 250 ml-Erlenmeyerkolben angeimpft und über Nacht (bis 16 h) bei 37°C unter Schütteln inkubiert.

Die frisch transformierten Stämme wurden außerdem auf festem Nährboden kultiviert. Die Kulturen werden auf LB-Amp-Agarplatten ausgestrichen und über Nacht bei 37°C inkubiert. Durch Verschließen der Petrischalen mit Parafilm und Lagern im Kühlraum (6°C) können die *E. coli*-Kulturen bis zur weiteren Benutzung gelagert werden (für etwa 4-6 Wochen).

Die Herstellung von *E. coli* Dauerkulturen erfolgte durch Animpfen von 1 ml LB-Amp-Glucose-Medium mit einer Einzelkolonie von einer Agarplatte, Inkubation über Nacht und anschließender Verdünnung eines 500 µl Aliquots dieser Kulturlösung mit 500 µl Glycerin (87%, autoklaviert). Bei –80°C sind diese Glycerol-Stocks über Jahre haltbar.[173]

5.17.2 Isolierung und Präzipitation genomischer DNA aus *Bacillus* Stämmen

Zur Isolation der genomischen DNA aus *Bacillus* Stämmen mittels Phenol-Chloroformextraktion wurde das Zellpellet aus 200 ml einer ÜN-Kultur in 20 ml Aufschlusspuffer I resuspendiert und 30 min lang bei 200 Upm und 37°C geschüttelt. Nach Zugaben von 2 ml SDS-Lösung (10% m/v) wurde weitere 60 min bei 37°C und 200 Upm inkubiert. Zur Entfernung von Zelltrümmern wurde zweimal im Verhältnis 1:1 mit einer Phenol/Chloroform-Lösung versetzt und 30 min bei 5000 Upm zentrifugiert. Die wässrige Phase wurde mit 0,1 Volumen Natriumacetat-Lösung (3 M; pH 4,8) und 0,6 Volumen Isopropanol versetzt und die ausgefallene DNA 30 min bei 5000 Upm abzentrifugiert. Danach wurde die DNA mit 70% Ethanol gewaschen, im Vakuum getrocknet (SpeedVac, 30°C, 10 min) und in 1,5 ml TE-Puffer aufgenommen.

Die Isolation der genomischen DNA aus *Streptomyces coelicolor* erfolgte nach Kutchma et al.,[174] die Isolation der genomischen DNA aus *Candida rugosa* erfolgte nach Cryer et al.[175] und die Isolationen der genomischen DNA aus *Aspergillus nidulans* erfolgte aus Sporen.

5.17.3 Polymerase-Kettenreaktionen (PCR)

5.17.3.1 Prinzip

Die PCR dient zur spezifischen Amplifikation einer Ausgangs-DNA (Matrize oder template), d.h. spezifische DNA-Sequenzen können *in vitro* in einem Reaktionszyklus mit hoher Ausbeute amplifiziert werden.[176-178] Dabei wird ein Zyklus von drei Reaktionsschritten bis zu dreißig Mal wiederholt. Der erste Schritt eines jeden Zyklus ist die Hitze-Denaturierung des Ausgangs-DNA-Doppelstrangs bei ca. 95°C. An die beiden entstandenen DNA-Einzelstränge lagert sich während des sogenannten Anlagerungsschritts (*Annealing*) bei 45-65°C ein kurzes komplementäres DNA-Fragment (Primer) an, das als Startpunkt für die DNA-Replikation dient. Im letzten Schritt (*Elongation*) werden durch Verlängerung der Primer am 3´-Ende mittels thermostabiler Polymerasen (hier: *Taq*-Polymerase) bei 72°C zwei neue Stränge synthetisiert, die bei den folgenden Zyklen wieder als template zur Verfügung stehen. Letztendlich wird das sich zwischen den Primern befindliche DNA-Fragment exponentiell vermehrt, weswegen nur wenige DNA-Moleküle als Ausgangsmaterial notwendig sind. Für eine erfolgreiche Amplifikation ist es erforderlich, die Bedingungen so zu wählen, dass die Fehlerrate beim Einbau der Nucleotide durch die Polymerase so niedrig wie möglich gehalten wird. Unter optimalen Bedingungen liegt die Fehlerrate bei 0,001 bis 0,02%.[179,180] Die Fehler sind dabei über das ganze Produkt verteilt, wobei am häufigsten ein Austausch von Cytosin

gegen Thymin auftritt. Deletionen oder Insertionen einer Base kommen nur äußerst selten (1 Mutation pro 50.000 Basen), komplexere Mutationen überhaupt nicht vor.

Für die Bestimmung einer DNA Sequenz eines bekanntes Peptids hat sich eine Variante etablieren können, die als Touchdown PCR bekannt ist.[181] Hierbei wird für jeden zweiten Zyklus die Annealing Temperatur um 1°C verringert. Die PCR wird unter Verwendung eines Satzes degenerierter Primer durchgeführt, die für den N- und C-terminalen Bereich des Peptids codieren und einen bekannten Abstand in der Sequenz zueinander besitzen. Hierbei entstehen mehrere PCR Produkte, das gewünschte Produkt kann jedoch über die Größe des Fragments identifiziert werden. Diese Methode wurde zur Bestimmung von Carboxylesterasefragmente aus genomischer DNA verwendet.

Die Polymerasekettenreaktion kann durch Abwandlungen auch zur Erzeugung von Mutanten genutzt werden. Für die Punktmutationen wurde nach dem Prinzip des Stratagene QuikChange® Kits[182] verfahren. Dazu werden sogenannte Punktmutationsprimer eingesetzt, die die gewünschte Mutation in ihrer Nukleotidsequenz bereits enthalten. Nach der Denaturierung der Plasmid-DNA, die als Matrize dient, hybridisieren die Primer mit den komplementären DNA-Strängen und die eingesetzte *Pfu*-Polymerase amplifiziert das gesamte Plasmid ausgehend von den eingesetzten Primern. Nach der PCR wird das Restriktionsenzym *Dpn* I zum Reaktionsgemisch gegeben. Im Gegensatz zur parentalen DNA ist die amplifizierte DNA nicht methyliert und wird daher von *Dpn* I nicht verdaut. Danach kann die amplifizierte DNA in kompetente Zellen transformiert werden. Die Punktmutation kann auch als Sättigungs-PCR mit sogenannten wobble-Primern durchgeführt werden. Hierbei wird ein Primergemisch verwendet, so dass an der zu mutierenden Stelle nicht eine ausgewählte, sondern alle 20 unterschiedlichen Aminosäuren auftreten können. Durch diese Methode können die Methoden rationales Design und gerichtete Evolution miteinander kombiniert werden.

5.17.3.2 PCR-Reaktion zur Isolierung der Carboxylesterasefragmente aus genomischer DNA

Standard-Ansatz (100 µl):	10xPuffer	10 µl
	$MgCl_2$-Lsg.	8 µl
	dNTP-Mix	8 µl
	Forward primer	2 µl
	Reverse primer	2 µl
	Ausgangs-DNA (2 µg/ml)	2,5 µl
	Taq-Polymerase	0,5 µl

	ddH_2O	67 µl

Tabelle 12: Temperaturprogramm für Touchdown PCR. Werte in Klammern wurden für GC reiche DNA verwendet

Reaktionsschritt	Temperatur [°C]	Zeit [min]	Zyklenzahl
Denaturierung	95	3	1
Denaturierung	95	1	
Anlagerung	60 (65) – 1 pro Zyklus	1	15
Elongation	72	1	
Denaturierung	95	1	
Anlagerung	45 (50) – 1 pro Zyklus	1	15
Elongation	72	1	
Synthese	72	7	1

5.17.3.3 Punktmutationen mittels QuikChange® Kit

Standard-Ansatz (50 µl):	10xPuffer	5 µl
	DMSO	5 µl
	dNTP-Mix	1 µl
	Forward primer	5 µl
	Reverse primer	5 µl
	Ausgangs-DNA (Midi)	2 µl
	Pfu-Polymerase	1 µl
	ddH_2O	26 µl

Tabelle 13: Temperaturprogramm für die *QuikChange*® PCR

Reaktionsschritt	Temperatur [°C]	Zeit [min]	Zyklenzahl
Denaturierung	95	2	1
Denaturierung	95	1	
Anlagerung	55	1	15

Elongation	68	10	
Synthese	68	10	1

5.17.4 DNA-Präzipitation

Um DNA zu isolieren oder in höheren Konzentrationen zu erhalten, wurde sie mit Ethanol oder Isopropanol gefällt. Unter leicht alkalischen Bedingungen löst sich DNA am besten. Sie wird nach der Fällung daher gewöhnlich in TE-Puffer, pH 7,8 aufgenommen. Da bei manchen Experimenten EDTA störend ist, wird die DNA gelegentlich auch in Tris-HCl gelöst (Verdau mit Restriktionsenzymen). Zur Sequenzierung vorgesehene DNA wurde in ddH_2O gelöst, da sich Puffersalze dabei störend auswirken.

5.17.4.1 Isopropanol-Fällung

Zu 1 Volumen DNA wurde das 0,7-fache Volumen an Isopropanol gegeben, 10 min bei Raumtemperatur inkubiert und dann 10-15 min zentrifugiert (14000 Upm). Das durchsichtige, glasige Pellet wurde im Vakuum getrocknet. Um den Trocknungsvorgang zeitlich abzukürzen, ist es ratsam das Pellet noch mit 70%igem Ethanol zu waschen, das leichter flüchtig ist und ausgefallene Salze beseitigt. Nach nochmaliger zehnminütiger Zentrifugation bei 14000 Upm wird die DNA im Vakuum getrocknet (SpeedVac, 30°C, 10 min) und in 30 – 50 µl Wasser oder geeignetem Puffer gelöst.

5.17.4.2 Ethanol-Fällung

Zu 1 Volumen DNA wurden ein Zehntel des Volumens an Natriumacetat-Lösung (3 M; pH 4,8) und das 2,5-fache Vol. an 99,8%igem kalten Ethanol gegeben. Die Probe wurde 20 Minuten in den Gefrierschrank (-80°C) gestellt und dann 20 Minuten bei 4°C zentrifugiert (14000 rpm). Der Überstand wurde vorsichtig abgegossen und das salzhaltige Pellet mit dem 2,5-fachen Volumen an 70%igem kalten Ethanol gewaschen. Die Probe wurde 10 min bei 4°C zentrifugiert (14000 rpm) und das Pellet nach Verwerfen des Überstands im Vakuum getrocknet (s. Isopropanolfällung). Danach wurde ebenfalls in 30-50 µl Wasser oder Puffer gelöst.

5.17.5 Transformation von Plasmid-DNA in *E. coli*

Unter Transformation versteht man Aufnahme freier, löslicher DNA aus dem Medium durch einen Rezipienten. Die natürliche Aufnahmebereitschaft für DNA (Kompetenz) ist für eine Vielzahl von Bakterien (*Acinetobacter, Bacillus, Pseudomonas* u.a.) beschrieben worden.[183] Bei *E. coli*, das hierfür kein System besitzt, können die Zellen durch eine geeignete Vorbehandlung ebenfalls transformationskompetent gemacht werden. Der Mechanismus des

Plasmid-Transfers in die *E. coli*-Zelle ist bislang noch nicht vollständig aufgeklärt, man geht von einer passiven Aufnahme, induziert durch Hitzeschock (42°C) oder Elektroporation, aus.

5.17.5.1 Transformation nach der TSS-Methode[184]

Um vollständige Plasmide in *E. coli* DH5α oder BL21(DE3) zu transformieren wurde die TSS-Methode verwendet: Aus einer ÜN-Kultur ohne Selektionsantibiotikum wurden 50 ml LB-Medium 1:50 angeimpft und bei 37°C und 200 Upm bis zu einer OD_{578} von 0,4 – 0,6 inkubiert. Die Zellen werden danach abzentrifugiert (4000 Upm, 10 min), in 2 ml TSS resuspendiert, auf Eis inkubiert und schließlich in 200 µl-Aliquots aufgeteilt und nach Schockgefrieren mit flüssigem Stickstoff bei –80°C gelagert. 200 µl der so kompetent gemachten Zellen wurden nach Vermischen mit 1 µl DNA 20 min auf Eis inkubiert, im Anschluß erfolgte der Hitzeschock, 45 s bei 42°C. Danach wurden 800 µl LB-Medium als Nährmedium hinzugegeben und 1 h bei 37°C und 200 Upm kultiviert. Die Zellen wurden 3 min bei 3000 Upm abzentrifugiert, der Überstand bis auf einen kleinen Rest verworfen, die Zellen resuspendiert und mit Hilfe eines Drigalski-Spatels auf einer LB-Amp- oder LB-Amp-Glucose-Agar-Platte ausgestrichen. Danach erfolgte die Inkubation bei 37°C über Nacht.

5.17.6 Plasmid-Präparation aus *E. coli*

5.17.6.1 Prinzip der Plasmid-Präparation

Die Plasmid-Präparationen wurden nach dem Prinzip der alkalischen Lyse durchgeführt. Die Zellen werden dabei in Puffer mit RNAse resuspendiert. Anschließend erfolgte die alkalische Lyse der Zellmembran durch SDS und damit verbunden die Denaturierung der Proteine sowie der DNA. Das darauffolgende Absenken des pH-Wertes bewirkt eine Renaturierung der Plasmid-DNA, während die chromosomale DNA ein Netzwerk bildet, das zusammen mit den Proteinen ein Präzipitat bildet, das sich durch Abzentrifugieren entfernen lässt. Die Plasmid-DNA bleibt in Lösung und kann so konzentriert und gereinigt werden.

5.17.6.2 Mini-Plasmid-Präparation mit dem QIAprep Spin Miniprep Kit

Für Anwendungen, die besonders reine Plasmid-DNA erfordern (PCR-Reaktionen, Sequenzierreaktionen) wurde die Mini-Plasmid-Präparationsmethode verwendet. Nach Abtrennung der Zellen einer 2-ml-ÜN-Kultur durch Zentrifugation (14000 Upm, 1 min) und Verwerfen des Kulturüberstandes erfolgte die weitere Plasmidreinigung nach Angaben des Herstellers Qiagen. Die DNA wurde danach in 30 – 50 µl 10 mM Tris-Puffer (pH 8,5) gelöst.

5.17.6.3 Midi-Plasmid-Präparation mit dem Plasmid Midi Kit

Zur Präparation größerer Mengen Plasmid-DNA wurde eine 50-ml-ÜN-Kultur 10 min bei 5000 Upm (4°C) zentrifugiert. Die Plasmid-Isolation erfolgte nach den Angaben des Herstellers Qiagen, wobei die Volumenangaben auf das Doppelte erhöht wurden. Gelöst wurde die DNA danach in 250 µl TE-Puffer.

5.17.6.4 Mini-Plasmid-Präparation (Schnelltest)

Für Anwendungen, die weder viel noch hochreine DNA erfordern (analytische Restriktionsverdauungen zur Kontrolle von Plasmiden) diente diese Methode, die abgewandelt nach den Anweisungen des FlexiPrep-Kits (Pharmacia) durchgeführt wurde.
Zur Zellabtrennung wurden 1,5 ml ÜN-Kultur in ein Eppendorf-Gefäß überführt, 30 sec bei 14000 Upm zentrifugiert und das entstandene Zellpellet so weit wie möglich vom Überstand befreit. Nach Zugabe von 200 µl Lsg. I und Vortexen wurden die nun resuspendierten Zellen durch Zugabe von 200 µl Lsg. II und vorsichtiges Schütteln lysiert. Die Zugabe der Lsg. III bewirkte unter vorsichtigem Schütteln das Ausflocken der Zellproteine und der genomischen DNA. Sie wurden durch Abzentrifugieren (14000 Upm, 5 min) entfernt. Die Plasmid-DNA im Überstand wurde durch Isopropanol-Fällung präzipitiert und in 15 µl TE-Puffer resuspendiert.

5.17.7 DNA-Sequenzierung

5.17.7.1 Theoretische Grundlagen

DNA wird nach der fluoreszenzmarkierten Cycle Sequencing Methode, abgeleitet von der Sanger-Didesoxy-Methode, sequenziert.[185] In einer „Dye-Terminator"-Sequenzierungsreaktion werden dabei DNA-Fragmente durch kontrollierten Abbruch der enzymatischen Replikation erzeugt. Thermostabile *Taq*-Polymerase wird verwendet, um eine bestimmte Sequenz einer einsträngigen DNA zu kopieren. Die Vervielfältigung wird durch ein kurzes komplementäres DNA-Fragment (Primer) gestartet. Zusätzlich zu den vier dNTPs enthält die Reaktionsmischung ihre 2,3-Didesoxy-Analoga. Ein Einbau dieser Analoga bewirkt den Abbruch der DNA-Replikation aufgrund der fehlenden 3´-OH-Gruppe, da der Aufbau einer Phosphodiesterbindung zu einem weiteren Nukleotid nicht mehr möglich ist. Es entstehen also Fragmente unterschiedlicher Länge, die am 3´-Ende ein Didesoxy-Analog enthalten. Die Didesoxy-Analoga sind mit vier verschiedenen Farbstoffen markiert, die nach Laseranregung fluoreszieren – je nach Base mit einer anderen Wellenlänge.[186,187]
Mit dieser Mischung von DNA-Fragmenten wird eine Gelelektrophorese durchgeführt. Die nach ihrer Größe getrennten Fragmente werden anhand der Fluoreszenz des endständigen

Didesoxy-Analogs detektiert.[188] Die Detektion bei dieser Vierfarbentechnik erfolgt durch zwei Argonlaser mit Emissionsbanden von 488 und 514 nm. Das senkrecht emitierte Fluoreszenslicht wird von Fotodioden hinter dem Gel gemessen und mit Hilfe eines angeschlossenen EDV-Systems ausgewertet.
Um lange DNA-Stränge zu sequenzieren, werden in verschiedenen Ansätzen mehrere Primer eingesetzt, die sich über den gesamten Strang verteilen. Die Genauigkeit der Sequenzierung wird erhöht, indem beide Stränge einer doppelsträngigen DNA sequenziert werden.

5.17.7.2 Probenvorbereitung

Sequenzierreaktion:
In einem Reaktionsvolumen von 20 µl wurden üblicherweise 4 µl Premix (BigDye Terminator Ready Reaction Kit) und 5 µl Primer (1 pmol/µl) zu 200 – 500 ng DNA (3 – 5 µl Mini-Plasmid-Präparation) gegeben.

Tabelle 14: Temperaturprogramm für die Sequenzier-PCR

Reaktionsschritt	Temperatur [°C]	Zeit [min]	Zyklenzahl
Denaturierung	95	4	1
Denaturierung	95	0,7	
Anlagerung	55	0,5	25
Elongation	60	4	
Synthese	60	4	1

Aufarbeitung:
Nach Durchführung einer Ethanolfällung wurden die Proben jeweils in 3 µl Formamid / 25 mM EDTA, pH 8 (5:1) mit einer Spatelspitze Dextran Blau aufgelöst, 2 min bei 95°C denaturiert und nach dem Abkühlen auf RT direkt auf das Sequenziergel aufgetragen (1 µl bei halber; 0,5 µl bei voller Beladung des Sequenziergels).

5.17.7.3 Gießen des Polyacrylamid-Gels und Durchführung

Beide Gelglasplatten wurden gründlich mit Alkanox® gewaschen und mit Wasser gespült. Danach wurde mit Isopropanol nachbehandelt und mit Druckluft getrocknet und ein Abstandshalter (Spacer) zwischen die Glasplatten gelegt. Für das 5,25%ige Page-Plus-Gel wurden der Harnstoff, TBE-Puffer, die 40%ige Page-Plus-Lösung und Wasser zusammengegeben. Der Harnstoff wurde gelöst, die Lösung anschließend filtriert (Millipore

Filter GVWP 0,22 µm) und entgast. Durch Zugabe von TEMED und APS-Lösung wurde die Polymerisation gestartet, die Lösung zum Aushärten zwischen die Glasplatten gegossen und der Vorkamm eingesetzt. Das Gel konnte nach zwei Stunden in den DNA-Sequenzierer eingespannt werden, wobei der Vorkamm durch einen „Haifisch-Kamm“ zur Probenaufgabe ausgetauscht wird. Nach Kontrolle des Basissignals erfolgte ein einstündiger Vorlauf. Die eigentliche Elektrophorese dauerte 12 h (Sequenzierbedingungen: 2,4 kV / 50 mA / 200 W / 50°C / 40 mV Laser Power).

5.18 Expression, Reinigung und Charakterisierung von Proteinen

Methoden zur Validierung des Anker-Konzepts wurden unter der Anleitung von Dr. Erik Henke und der Mithilfe der wissenschaftlichen Hilfskraft Nadine Pollak durchgeführt.

5.18.1 Expression des Zielproteins im Schüttelkolben

500 ml LB-Amp-Medium wurden mit 500 µl einer ÜN-Kultur der rekombinanten Zellen inokuliert. Man inkubierte im Schüttelkolben bei 30 °C bis zu einer OD_{600} von 0,8-1,0 und induzierte die Expression des Zielproteins unter dem hitzeabhängigen λ-Promoter durch die Erhöhung der Temperatur auf 42 °C für 4 h. Danach wurden die Zellen abzentrifugiert (3000 g, 20 min, 4 °C). Um die Reste des Mediums zu entfernen, wurde das Pellet in 50 ml Reinigungspuffer resuspendiert, nochmals abzentrifugiert und der Überstand verworfen. Nach erneutem Aufnehmen in 20 ml Reinigungspuffer wurden die Zellen unter Eiskühlung mittels Ultraschall aufgeschlossen. Zelltrümmer wurden durch Zentrifugation (8000 g, 30 min, 4 °C) entfernt.

5.18.2 Gewinnung der aktiven BCL durch Chaperon-vermittelte Faltung

Die über die heterologe Expression in *Escherichia coli* erhaltene BCL liegt in *inclusion bodies* vor, da für die native Faltung der BCL ein Chaperon aus *Burkholderia cepacia* benötigt wird. Zur Gewinnung der aktiven BCL und deren Varianten, mussten diese deshalb in Gegenwart des Chaperons umgefaltet werden. Dies erfolgte nach der von Quyen et al. beschriebenen Methode.[172]

Das die Lipase enthaltende *E. coli* Zellextrakt wurde direkt zur Umfaltung eingesetzt. In dH_2O wurden äquimolare Mengen Lipase und Chaperon für 48 h bei 8 °C inkubiert (Verdünnung von 1:1000 der Lipase und des Chaperons). Die umgefaltete, aktive Lipase wurde über Crossflow Filtration (10 kDa Membran) und anschließender Ultrafiltration (10

kDa Membran) aufkonzentriert. Das Konzentrat wurde in Petri-Schalen gegossen, bei -80 °C ca. eine Stunde eingefroren und lyophilisiert.

5.18.3 SDS-Gel-Elektrophorese (SDS-PAGE)

5.18.3.1 Theoretische Grundlagen

Proteine werden bei der SDS-PAGE überwiegend nach ihrer Masse unter denaturierenden Bedingungen aufgetrennt.[189] Ein Proteingemisch wird zunächst in einer Natriumdodecylsulfat-Lösung (SDS) gelöst, einem anionischen Detergenz, das nichtkovalente Wechselwirkungen unter nativen Proteinen aufhebt. Außerdem wird Mercaptoethanol zur Disulfidbrückenreduktion zugegeben. Dodecylsulfat-Anionen binden an die Proteinkette und bilden mit dem denaturierten Protein einen Komplex, dessen negative Ladung ungefähr proportional zur Masse des Proteins ist. Die SDS-Proteinkomplexe verschiedener Proteine unterscheiden sich damit nur noch in ihrer Größe (Molekulargewicht bzw. Stokes-Radius) und haben vergleichbare hydrodynamische Eigenschaften. Mit diesen Komplexen wird dann eine vertikale Polyacrylamid-Gel-Elektrophorese durchgeführt. Die Wanderungsgeschwindigkeit hängt von der Ladung des SDS-Proteinkomplexes, der Porengröße des Gels und der angelegten Spannung ab. Schließlich werden die Proteine im Gel durch einen Farbstoff (Coomassie Blue) sichtbar gemacht.

5.18.3.2 Durchführung

Gießen des Gels:

Zunächst wurde das Trenngel gegossen. Dazu wurden 2 ml Lower Tris; 3,33 ml Acrylamid-Lösung (30%); 2,67 ml dH_2O und 40 µl APS-Lösung zusammenpipettiert. Mit Zugabe von 4 µl TEMED wurde der Polymerisationsprozeß gestartet und das Gel sofort bis etwa 2 cm unterhalb der oberen Gelbegrenzung zwischen die vorbereiteten Glasplatten gegossen und dann mit Isopropanol überschichtet, um eine gerade Geloberfläche zu erhalten. Nach einer Polymerisationszeit von 30 min wurde das Isopropanol abgegossen und Reste durch Spülen mit dH_2O entfernt. Im Anschluß wurde das Sammelgel gegossen. Dazu wurden 1 ml Upper Tris; 0,52 ml Acrylamid-Lösung (30%); 2,47 ml dH_2O und 40 µl APS-Lösung zusammenpipettiert. Mit Zugabe von 4 µl TEMED wurde der Polymerisationsprozeß gestartet und das Gel sofort auf das polymerisierte Trenngel gegossen. Danach wurde der Probenkamm luftblasenfrei zwischen die Glasplatten gesetzt. Nach weiteren 30 min konnte der Probenkamm vorsichtig aus dem Sammelgel entfernt und das Gel ebenfalls luftblasenfrei in die Elektrophoresekammer eingesetzt werden.

Probenvorbereitung:
Bei Roh- oder Reinextrakten wurden je 10 µl der Proben 1:1 mit SDS-Probenpuffer gemischt, für 10 min bei 95°C gekocht und nach dem Abkühlen auf das Gel aufgetragen. Bei Zellpellets, die durch Abzentrifugieren von 2 ml Kulturmedium gewonnen wurden, erfolgte die Vorbereitung durch Aufschlämmen des Pellets in 100 µl SDS-Probenpuffer und zehnminütigem Kochen bei 95°C.

Elektrophorese:
Die Elektrodenkammern wurden nach Einsetzen des Gels mit Elektrodenpuffer gefüllt und die Proben aufgetragen. Die Elektrophorese erfolgte für 12 min mit einer Stromstärke von 10 mA und anschließend etwa für 60 min mit 25 mA, bis die Lauffront das Ende des Gels erreichte. Als Proteinstandard wurden 10 µl eines Breitbandstandards der Firma BIO-RAD aufgetragen.

Färben und Trocknen des Gels:
Das SDS-Gel wurde nach der Elektrophorese in der Coomassie-Lösung mindestens 1 h gefärbt. Danach wurde das Gel zwei Stunden in die Entfärbelösung gelegt, die nach einer halben Stunde gewechselt wurde. Zur Aufbewahrung wurde das Gel im Geltrockner zwischen einer Cellophan-Folie und einem Cellulose-Filterpapier (leichtes Vakuum, 80°C) 2 h getrocknet.

5.18.4 Bestimmung des Proteingehalts mittels BCA-Nachweis

Proteinkonzentrationen wurden mit Hilfe des BCA-Proteinnachweissystems spektrometrisch bei 562 nm nach Herstellerangaben entsprechend dem Standardprotokoll bestimmt.
Kalibiriergeraden wurden mit BSA-Verdünnungen aufgenommen.

5.19 Analytische Methoden

Methoden zur Validierung des Anker-Konzepts wurden unter der Anleitung von Dr. Erik Henke und der Mithilfe der wissenschaftlichen Hilfskraft Nadine Pollak durchgeführt.

5.19.1 Aktivitätsbestimmung von Lipasen mittels *p*-Nitrophenylpalmitat (photometrischer Assay)

Zur schellen Bestimmung der Aktivität von Lipasepräparationen wurde ein photometrischer Assay verwendet. Grundlage ist die Hydrolyse von *p*-Nitrophenylpalmitat durch die Lipase.

Das dabei freigesetzte *p*-Nitrophenolat ($\varepsilon_{410nm,\ pH7,5}$ = 14500) wird photometrisch durch Absorption bei einer Wellenlänge von 410 nm detektiert.

Die Messungen wurden in einem UV/VIS-Spektralphotometer mit 6-fach Wechselprobenhalter bei Raumtemperatur durchgeführt.

In einer 1 ml-Mikroküvette wurden 100 µl Lipase in TrisHCL-Puffer (pH 7,5, 100 mM) mit 0,8% (w/v) Cholat und 1% (w/v) Gummi Arabicum vorgelegt und mit 800 µl der Pufferlösung verdünnt. Als Referenz diente eine Küvette mit 900 µl Pufferlösung. Nach Aufnahme der Hintergrundabsorption wurde die Reaktion durch Zugabe von *p*-Nitrophenylpalmitat (10 mM in Isopropanol) gestartet. Die Aufnahme der Hydrolysegeschwindigkeit erfolgte durch Absorptionsmessungen im Abstand von 10 sek. über 3 min. Aus der Änderung der Absorption über die Zeit konnte die Geschwindigkeit der Umsetzung berechnet werden.

5.19.1 Aktivitätsbestimmung von Lipasen am pH-Stat

In demineralisiertes Wasser wurden der Substratester (Endkonzentration 20 mM) und 2% (w/v) Gummi Arabicum gegeben. Die Mischung wurde im Ultra-turrax homogenisiert. Nach der Homogenisierung wurde $CaCl_2$ zu einer Endkonzentration von 20 mM zugesetzt. In der thermostatisierten Reaktionskammer des ph-Staten wurden 20 ml der Emulsion vorgelegt, und nach Einstellen des pH-Werts die Lipase zugesetzt.

Die Titration der freigesetzten Fettsäure erfolgte automatisch, der Verbrauch an Natronlauge wurde über einen Schreiber festgehalten. Aus dem Verbrauch pro Zeiteinheit ließ sich die Aktivität berechnen, wobei eine Einheit (U) als die Menge Lipase definiert wurde, die ein µl Substrat in einer Minute freisetzt.

6 Literatur

1. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol 1965;8(2):357-366.

2. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. Embo J 1986;5(4):823-826.

3. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167-339.

4. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372(6507):631-634.

5. 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.

6. Holmquist M. Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms. Curr Protein Pept Sci 2000;1(2):209-235.

7. Bray JE, Todd AE, Pearl FM, Thornton JM, Orengo CA. The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. Protein Eng 2000;13(3):153-165.

8. Salem GM, Hutchinson EG, Orengo CA, Thornton JM. Correlation of observed fold frequency with the occurrence of local structural motifs. J Mol Biol 1999;287(5):969-981.

9. Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951-960.

10. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.6. Seattle: Distributed by author. Department of Genome Sciences, University of Washington; 2004.

11. Morant M, Hehn A, Werck-Reichhart D. Conservation and diversity of gene families explored using the CODEHOP strategy in higher plants. BMC Plant Biol 2002;2(1):7.

12. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291(5507):1304-1351.

13. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC. A whole-genome assembly of Drosophila. Science 2000;287(5461):2196-2204.

14. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269(5223):496-512.

15. Gaasterland T. Structural genomics taking shape. Trends Genet 1998;14(4):135.

16. Kim SH. Shining a light on structural genomics. Nat Struct Biol 1998;5 Suppl:643-645.

17. May AC. Optimal classification of protein sequences and selection of representative sets from multiple alignments: application to homologous families and lessons for structural genomics. Protein Eng 2001;14(4):209-217.

18. Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A. Recent improvements to the PROSITE database. Nucleic Acids Res 2004;32 Database issue:D134-137.

19. Parry-Smith DJ, Attwood TK. ADSP--a new package for computational sequence analysis. Comput Appl Biosci 1992;8(5):451-459.

20. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 1995;163(2):GC17-26.

21. Henikoff S, Henikoff JG, Pietrokovski S. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics 1999;15(6):471-479.

22. Attwood TK. The PRINTS database: a resource for identification of protein families. Brief Bioinform 2002;3(3):252-263.

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

24. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994;235(5):1501-1531.

25. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32 Database issue:D138-141.

26. Srinivasarao GY, Yeh LS, Marzec CR, Orcutt BC, Barker WC, Pfeiffer F. Database of protein sequence alignments: PIR-ALN. Nucleic Acids Res 1999;27(1):284-285.

27. Murvai J, Vlahovicek K, Barta E, Cataletto B, Pongor S. The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments. Nucleic Acids Res 2000;28(1):260-262.

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

29. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85(8):2444-2448.

30. 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):3389-3402.

31. Cousin X, Hotelier T, Giles K, Lievin P, Toutant JP, Chatonnet A. The alpha/beta fold family of proteins database and the cholinesterase gene server ESTHER. Nucleic Acids Res 1997;25(1):143-146.

32. Hotelier T, Renault L, Cousin X, Negre V, Marchot P, Chatonnet A. ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins. Nucleic Acids Res 2004;32 Database issue:D145-147.

33. Coutinho P, Henrissat B. Carbohydrate-active enzymes: an integrated database approach. In: H.J. Gilbert GD, B. Henrissat and B. Svensson, editor. Recent Advances in Carbohydrate Bioengineering. Cambridge: The Royal Society of Chemistry; 1999. p 3-12.

34. Dayhoff M, Eck R. Atlas of Protein Sequence and Structure. Silver Springs, Maryland: NBRF Press; 1966.

35. 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.

36. Waterman MS, Smith TF, Beyer WA. Some Biological Sequence Metrics. Adv Math 1976;20(3):367-387.

37. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147(1):195-197.

38. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89(22):10915-10919.

39. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326(6111):347-352.

40. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2(2):171-178.

41. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257(2):342-358.

42. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286(5438):295-299.

43. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 1999;287(1):187-198.

44. Ptitsyn OB. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? J Mol Biol 1998;278(3):655-666.

45. Ptitsyn OB, Ting KL. Non-functional conserved residues in globins and their possible role as a folding nucleus. J Mol Biol 1999;291(3):671-682.

46. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291(1):177-196.

47. Dubchak I, Frazer K. Multi-species sequence comparison: the next frontier in genome annotation. Genome Biol 2003;4(12):122.

48. Moore G. Cramming more components onto integrated circuits. Electronics 1965;38:114-117.

49. Rost B. Marrying structure and genomics. Structure 1998;6(3):259-263.

50. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25-29.

51. Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Methods Enzymol 1996;266:114-128.

52. Fujibuchi W, Goto S, Migimatsu H, Uchiyama I, Ogiwara A, Akiyama Y, Kanehisa M. DBGET/LinkDB: an integrated database retrieval system. Pac Symp Biocomput 1998:683-694.

53. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. Methods Enzymol 1996;266:141-162.

54. Davidson SB, Overton C, Buneman P. Challenges in integrating biological data sources. J Comput Biol 1995;2(4):557-572.

55. Sheth AP, Larson JA. Federated Database-Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. Comput Surv 1990;22(3):183-236.

56. Davidson SB, Crabtree J, Brunk BP, Schug J, Tannen V, Overton GC, Stoeckert CJ. K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. Ibm Syst J 2001;40(2):512-531.

57. Inmon WH. Building the Data Warehouse. New York: John Wiley & Sons; 2002.

58. Cattell RGG, Barry DK. The object database standard, ODMG 2.0. San Francisco, Calif.: Morgan Kaufmann Publishers; 1997. 270 p. p.

59. Codd EF. A Relational Model of Data for Large Shared Data Banks. Commun Acm 1970;13(6):377-&.

60. Codd EF. Relational Database - a Practical Foundation for Productivity. Commun Acm 1982;25(2):109-117.

61. Blow DM, Birktoft JJ, Hartley BS. Role of a buried acid group in the mechanism of action of chymotrypsin. Nature 1969;221(178):337-340.

62. Wright CS, Alden RA, Kraut J. Structure of subtilisin BPN' at 2.5 angstrom resolution. Nature 1969;221(177):235-242.

63. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32(1):D226-229.

64. Pathak D, Ollis D. Refined structure of dienelactone hydrolase at 1.8 A. J Mol Biol 1990;214(2):497-525.

65. Franken SM, Rozeboom HJ, Kalk KH, Dijkstra BW. Crystal structure of haloalkane dehalogenase: an enzyme to detoxify halogenated alkanes. Embo J 1991;10(6):1297-1302.

66. Liao DI, Remington SJ. Structure of wheat serine carboxypeptidase II at 3.5-A resolution. A new class of serine proteinase. J Biol Chem 1990;265(12):6528-6531.

67. Sussman JL, Harel M, Frolow F, Oefner C, Goldman A, Toker L, Silman I. Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein. Science 1991;253(5022):872-879.

68. Schrag JD, Li YG, Wu S, Cygler M. Ser-His-Glu triad forms the catalytic site of the lipase from Geotrichum candidum. Nature 1991;351(6329):761-764.

69. Slabas AR, Windust J, Sidebottom CM. Does sequence similarity of human choline esterase, Torpedo acetylcholine esterase and Geotrichum candidum lipase reveal the active site serine residue? Biochem J 1990;269(1):279-280.

70. Ding J, McGrath WJ, Sweet RM, Mangel WF. Crystal structure of the human adenovirus proteinase with its 11 amino acid cofactor. Embo J 1996;15(8):1778-1783.

71. Higaki JN, Evnin LB, Craik CS. Introduction of a cysteine protease active site into trypsin. Biochemistry 1989;28(24):9256-9263.

72. Clark PI, Lowe G. Conversion of the active-site cysteine residue of papain into a dehydro-serine, a serine and a glycine residue. Eur J Biochem 1978;84(1):293-299.

73. Brenner S. The molecular evolution of genes and proteins: a tale of two serines. Nature 1988;334(6182):528-530.

74. Lawson DM, Derewenda U, Serre L, Ferri S, Szittner R, Wei Y, Meighen EA, Derewenda ZS. Structure of a myristoyl-ACP-specific thioesterase from Vibrio harveyi. Biochemistry 1994;33(32):9382-9388.

75. Schrag JD, Vernet T, Laramee L, Thomas DY, Recktenwald A, Okoniewska M, Ziomek E, Cygler M. Redesigning the active site of Geotrichum candidum lipase. Protein Eng 1995;8(8):835-842.

76. Rogalska E, Cudrey C, Ferrato F, Verger R. Stereoselective hydrolysis of triglycerides by animal and microbial lipases. Chirality 1993;5(1):24-30.

77. Schmid RD, Verger R. Lipases: Interfacial enzymes with attractive applications. Angew Chem Int Edit 1998;37(12):1609-1633.

78. Sarda L, Desnuelle P. [Actions of pancreatic lipase on esters in emulsions.]. Biochim Biophys Acta 1958;30(3):513-521.

79. Desnuelle P, Sarda L, Ailhaud G. Inhibition De La Lipase Pancreatique Par Le Diethyl-P-Nitrophenyl Phosphate En Emulsion. Biochimica Et Biophysica Acta 1960;37(3):570-571.

80. Hjorth A, Carriere F, Cudrey C, Woldike H, Boel E, Lawson DM, Ferrato F, Cambillau C, Dodson GG, Thim L, Verger R. A Structural Domain (the Lid) Found in Pancreatic Lipases Is Absent in the Guinea-Pig (Phospho)Lipase. Biochemistry 1993;32(18):4702-4707.

81. Lesuisse E, Schanck K, Colson C. Purification and Preliminary Characterization of the Extracellular Lipase of Bacillus-Subtilis 168, an Extremely Basic Ph-Tolerant Enzyme. European Journal of Biochemistry 1993;216(1):155-160.

82. Martinez C, De Geus P, Lauwereys M, Matthyssens G, Cambillau C. Fusarium solani cutinase is a lipolytic enzyme with a catalytic serine accessible to solvent. Nature 1992;356(6370):615-618.

83. Derewenda U, Brzozowski AM, Lawson DM, Derewenda ZS. Catalysis at the interface: the anatomy of a conformational change in a triglyceride lipase. Biochemistry 1992;31(5):1532-1541.

84. Grochulski P, Li Y, Schrag JD, Cygler M. Two conformational states of Candida rugosa lipase. Protein Sci 1994;3(1):82-91.

85. van Tilbeurgh H, Egloff MP, Martinez C, Rugani N, Verger R, Cambillau C. Interfacial activation of the lipase-procolipase complex by mixed micelles revealed by X-ray crystallography. Nature 1993;362(6423):814-820.

86. Robertus JD, Kraut J, Alden RA, Birktoft JJ. Subtilisin; a stereochemical mechanism involving transition-state stabilization. Biochemistry 1972;11(23):4293-4303.

87. Carter P, Wells JA. Dissecting the catalytic triad of a serine protease. Nature 1988;332(6164):564-568.

88. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31(1):365-370.

89. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.

90. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17(4):355-362.

91. Krause A, Stoye J, Vingron M. The SYSTERS protein sequence cluster set. Nucleic Acids Res 2000;28(1):270-272.

92. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. Nucleic Acids Res 2004;32 Database issue:D23-26.

93. Schwede T, Diemand A, Guex N, Peitsch MC. Protein structure computing in the genomic era. Res Microbiol 2000;151(2):107-112.

94. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577-2637.

95. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22(22):4673-4680.

96. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 2002;18(3):502-504.

97. McLachlan AD. Rapid Comparison of Protein Structres. Acta Cryst 1992;A38:871-873.

98. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000;16(6):276-277.

99. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 2001;29(1):221-222.

100. Uppenberg J, Ohrner N, Norin M, Hult K, Kleywegt GJ, Patkar S, Waagen V, Anthonsen T, Jones TA. Crystallographic and molecular-modeling studies of lipase B from Candida antarctica reveal a stereospecificity pocket for secondary alcohols. Biochemistry 1995;34(51):16838-16851.

101. Schrag JD, Li Y, Cygler M, Lang D, Burgdorf T, Hecht HJ, Schmid R, Schomburg D, Rydel TJ, Oliver JD, Strickland LC, Dunaway CM, Larson SB, Day J, McPherson A. The open conformation of a Pseudomonas lipase. Structure 1997;5(2):187-202.

102. Egloff MP, Marguet F, Buono G, Verger R, Cambillau C, van Tilbeurgh H. The 2.46 A resolution structure of the pancreatic lipase-colipase complex inhibited by a C11 alkyl phosphonate. Biochemistry 1995;34(9):2751-2762.

103. Brzozowski AM, Derewenda U, Derewenda ZS, Dodson GG, Lawson DM, Turkenburg JP, Bjorkling F, Huge-Jensen B, Patkar SA, Thim L. A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. Nature 1991;351(6326):491-494.

104. Grochulski P, Bouthillier F, Kazlauskas RJ, Serreqi AN, Schrag JD, Ziomek E, Cygler M. Analogs of reaction intermediates identify a unique substrate binding site in Candida rugosa lipase. Biochemistry 1994;33(12):3494-3500.

105. Longhi S, Nicolas A, Creveld L, Egmond M, Verrips CT, de Vlieg J, Martinez C, Cambillau C. Dynamics of Fusarium solani cutinase investigated through structural comparison among different crystal forms of its variants. Proteins 1996;26(4):442-458.

106. Harel M, Quinn DM, Nair HK, Silman I, Sussman JL. The X-ray structure of a transition state analog complex reveals the molecular origins of the catalytic power and substrate specificity of acetylcholinesterase. J Am Chem Soc 1996;118(10):2340-2346.

107. Hecht HJ, Sobek H, Haag T, Pfeifer O, van Pee KH. The metal-ion-free oxidoreductase from Streptomyces aureofaciens has an alpha/beta hydrolase fold. Nat Struct Biol 1994;1(8):532-537.

108. Zock J, Cantwell C, Swartling J, Hodges R, Pohl T, Sutton K, Rosteck P, Jr., McGilvray D, Queener S. The Bacillus subtilis pnbA gene encoding p-nitrobenzyl esterase: cloning, sequence and high-level expression in Escherichia coli. Gene 1994;151(1-2):37-43.

109. Alberghina L, Lotti M. Cloning, sequencing, and expression of Candida rugosa lipases. Methods Enzymol 1997;284:246-260.

110. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 1997;25(1):31-36.

111. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. Brief Bioinform 2002;3(3):246-251.

112. Zdobnov EM, Lopez R, Apweiler R, Etzold T. The EBI SRS server--recent developments. Bioinformatics 2002;18(2):368-373.

113. Arpigny JL, Jaeger KE. Bacterial lipolytic enzymes: classification and properties. Biochem J 1999;343 Pt 1:177-183.

114. Beer HD, Wohlfahrt G, McCarthy JE, Schomburg D, Schmid RD. Analysis of the catalytic mechanism of a fungal lipase using computer-aided design and structural mutants. Protein Eng 1996;9(6):507-517.

115. Holm L, Sander C. The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res 1996;24(1):206-209.

116. Lourenco PM, Almeida T, Mendonca D, Simoes F, Novo C. Searching for nitrile hydratase using the Consensus-Degenerate Hybrid Oligonucleotide Primers strategy. J Basic Microbiol 2004;44(3):203-214.

117. Williamson N, Brian P, Wellington EM. Molecular detection of bacterial and streptomycete chitinases in the environment. Antonie Van Leeuwenhoek 2000;78(3-4):315-321.

118. Bell PJ, Sunna A, Gibbs MD, Curach NC, Nevalainen H, Bergquist PL. Prospecting for novel lipase genes using PCR. Microbiology 2002;148(Pt 8):2283-2291.

119. Henke E, Pleiss J, Bornscheuer UT. Activity of lipases and esterases towards tertiary alcohols: insights into structure-function relationships. Angew Chem Int Ed Engl 2002;41(17):3211-3213.

120. Pleiss J, Mionetto N, Schmid RD. Probing the acyl binding site of acetylcholinesterase by protein engineering. J Mol Catal B: Enzym 1999;6:287-296.

121. Atomi H, Bornscheuer UT, Soumanou MM, Beer HD, Wohlfahrt G, Schmid RD. Microbial lipases-from screening to design. In Oils-Fats-Lipids, Proceeding of the Twenty-first World Congress of the International Society on Fat Research. 1996; Bridgewater. PJ Barnes and Associates. p 49-50.

122. Joerger RD, Haas MJ. Alteration of chain length selectivity of a Rhizopus delemar lipase through site-directed mutagenesis. Lipids 1994;29(6):377-384.

123. Klein RR, King G, Moreau RA, Haas MJ. Altered acyl chain length specificity of Rhizopus delemar lipase through mutagenesis and molecular modeling. Lipids 1997;32(2):123-130.

124. Halling PJ. Thermodynamic predictions for biocatalysis in nonconventional media: theory, tests, and recommendations for experimental design and analysis. Enzyme Microb Technol 1994;16(3):178-206.

125. Adlercreutz P. On the importance of the support material for enzymatic synthesis in organic media. Support effects at controlled water activity. Eur J Biochem 1991;199(3):609-614.

126. Verger R. Interfacial activation' of lipases: facts and artifacts. Trends Biotechnol 1997;15(1):32-38.

127. Rangheard MS, Langrand G, Triantaphylides C, Baratti J. Multi-competitive enzymatic reactions in organic media: a simple test for the determination of lipase fatty acid specificity. Biochim Biophys Acta 1989;1004(1):20-28.

128. Kirk O, Bjoerkling F, Godtfredsen SE, Larsen TO. Fatty acid specificity in lipase-catalyzed synthesis of glucoside esters. Biocatalysis 1992;6(2):127-134.

129. Berger M, Schneider MP. Lipases in organic solvents: The fatty acid chain length profile. Biotechnol Lett 1991;13:641-645.

130. Kraut J. Serine proteases: structure and mechanism of catalysis. Annu Rev Biochem 1977;46:331-358.

131. Derewenda U, Swenson L, Wei Y, Green R, Kobos PM, Joerger R, Haas MJ, Derewenda ZS. Conformational lability of lipases observed in the absence of an oil-water interface: crystallographic studies of enzymes from the fungi Humicola lanuginosa and Rhizopus delemar. J Lipid Res 1994;35(3):524-534.

132. Yamaguchi S, Mase T, Takeuchi K. Secretion of mono- and diacylglycerol lipase from Penicillium camembertii U-150 by Saccharomyces cerevisiae and site-directed mutagenesis of the putative catalytic sites of the lipase. Biosci Biotechnol Biochem 1992;56(2):315-319.

133. Nagao T, Shimada Y, Sugihara A, Tominaga Y. Cloning and nucleotide sequence of cDNA encoding a lipase from Fusarium heterosporum. J Biochem (Tokyo) 1994;116(3):536-540.

134. Kohno M, Funatsu J, Mikami B, Kugimiya W, Matsuo T, Morita Y. The crystal structure of lipase II from Rhizopus niveus at 2.2 A resolution. J Biochem (Tokyo) 1996;120(3):505-510.

135. Nicolas A, Egmond M, Verrips CT, de Vlieg J, Longhi S, Cambillau C, Martinez C. Contribution of cutinase serine 42 side chain to the stabilization of the oxyanion transition state. Biochemistry 1996;35(2):398-410.

136. Winkler FK, D'Arcy A, Hunziker W. Structure of human pancreatic lipase. Nature 1990;343(6260):771-774.

137. van Tilbeurgh H, Gargouri Y, Dezan C, Egloff MP, Nesa MP, Ruganie N, Sarda L, Verger R, Cambillau C. Crystallization of pancreatic procolipase and of its complex with pancreatic lipase. J Mol Biol 1993;229(2):552-554.

138. Lombardo D, Chapus C, Bourne Y, Cambillau C. Crystallization and preliminary X-ray study of horse pancreatic lipase. J Mol Biol 1989;205(1):259-261.

139. Withers-Martinez C, Carriere F, Verger R, Bourgeois D, Cambillau C. A pancreatic lipase with a phospholipase A1 activity: crystal structure of a chimeric pancreatic lipase-related protein 2 from guinea pig. Structure 1996;4(11):1363-1374.

140. Noble ME, Cleasby A, Johnson LN, Egmond MR, Frenken LG. The crystal structure of triacylglycerol lipase from Pseudomonas glumae reveals a partially redundant catalytic aspartate. FEBS Lett 1993;331(1-2):123-128.

141. Lang D, Hofmann B, Haalck L, Hecht HJ, Spener F, Schmid RD, Schomburg D. Crystal structure of a bacterial lipase from Chromobacterium viscosum ATCC 6918 refined at 1.6 angstroms resolution. J Mol Biol 1996;259(4):704-717.

142. Roussel A, Miled N, Berti-Dupuis L, Riviere M, Spinelli S, Berna P, Gruber V, Verger R, Cambillau C. Crystal structure of the open form of dog gastric lipase in complex with a phosphonate inhibitor. J Biol Chem 2002;277(3):2266-2274.

143. Yang J, Koga Y, Nakano H, Yamane T. Modifying the chain-length selectivity of the lipase from Burkholderia cepacia KWI-56 through in vitro combinatorial mutagenesis in the substrate-binding site. Protein Eng 2002;15(2):147-152.

144. Koga Y, Kato K, Nakano H, Yamane T. Inverting enantioselectivity of Burkholderia cepacia KWI-56 lipase by combinatorial mutation and high-throughput screening using single-molecule PCR and in vitro expression. J Mol Biol 2003;331(3):585-592.

145. Dall'Acqua W, Carter P. Substrate-assisted catalysis: molecular basis and biological significance. Protein Sci 2000;9(1):1-9.

146. Herrgard S, Gibas CJ, Subramaniam S. Role of an electrostatic network of residues in the enzymatic action of the Rhizomucor miehei lipase family. Biochemistry 2000;39(11):2921-2930.

147. Friedberg I, Margalit H. Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. Protein Sci 2002;11(2):350-360.

148. Fersht AR. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc Natl Acad Sci U S A 1995;92(24):10869-10873.

149. Hutchinson EG, Thornton JM. PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci 1996;5(2):212-220.

150. Shakhnovich EI. Folding nucleus: specific or multiple? Insights from lattice models and experiments. Fold Des 1998;3(6):R108-111; discussion R107.

151. Abkevich VI, Gutin AM, Shakhnovich EI. Specific nucleus as the transition state for protein folding: evidence from the lattice model. Biochemistry 1994;33(33):10026-10036.

152. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4(4):406-425.

153. Strimmer K, vonHaeseler A. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Molecular Biology and Evolution 1996;13(7):964-969.

154. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 2001;18(5):691-699.

155. Margush T, Mcmorris FR. Consensus N-Trees. B Math Biol 1981;43(2):239-244.

156. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics 2002;18(1):77-82.

157. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17(3):282-283.

158. Hubbard SJ, Thornton JM. NACCESS: Department of Biochemistry and Molecular Biology, University College London; 1993.

159. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55(3):379-400.

160. Hubbard SJ, Campbell SF, Thornton JM. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J Mol Biol 1991;220(2):507-530.

161. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196(3):641-656.

162. Churchill GA. Stochastic models for heterogeneous DNA sequences. Bull Math Biol 1989;51(1):79-94.

163. Lyngso RB, Pedersen CN, Nielsen H. Metrics and similarity measures for hidden Markov models. Proc Int Conf Intell Syst Mol Biol 1999:178-186.

164. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17(8):700-712.

165. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221(4612):709-713.

166. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28(7):849-857.

167. Smith RF, Smith TF. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc Natl Acad Sci U S A 1990;87(1):118-122.

168. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci 1995;4(8):1618-1632.

169. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 1993;262(5131):208-214.

170. Rose TM, Schultz ER, Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res 1998;26(7):1628-1635.

171. Belev TN, Singh M, McCarthy JE. A fully modular vector system for the optimization of gene expression in Escherichia coli. Plasmid 1991;26(2):147-150.

172. Quyen DT, Schmidt-Dannert C, Schmid RD. High-level formation of active Pseudomonas cepacia lipase after heterologous expression of the encoding gene and its modified chaperone in Escherichia coli and rapid in vitro refolding. Appl Environ Microbiol 1999;65(2):787-794.

173. Sambook J, Russell DW. Molecular Cloning. New York: Cold Spring Harbor Laboratory Press; 2001.

174. Kutchma AJ, Roberts MA, Knaebel DB, Crawford DL. Small-scale isolation of genomic DNA from Streptomyces mycelia or spores. Biotechniques 1998;24(3):452-456.

175. Cryer DR, Eccleshall R, Marmur J. Isolation of yeast DNA. Methods Cell Biol 1975;12:39-44.

176. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 1985;230(4732):1350-1354.

177. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988;239(4839):487-491.

178. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 1986;51 Pt 1:263-273.

179. Tindall KR, Kunkel TA. Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. Biochemistry 1988;27(16):6008-6013.

180. Cadwell RC, Joyce GF. Randomization of genes by PCR mutagenesis. PCR Methods Appl 1992;2(1):28-33.

181. Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS. 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res 1991;19(14):4008.

182. Kunkel TA. Rapid and efficient site-specific mutagenesis without phenotypic selection. Proc Natl Acad Sci U S A 1985;82(2):488-492.

183. Schlegel HG. Allgemeine Mikrobiologie, 7. Aufl. Stuttgart: Georg Thieme Verlag; 1992.

184. Chung CT, Niemela SL, Miller RH. One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. Proc Natl Acad Sci U S A 1989;86(7):2172-2175.

185. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74(12):5463-5467.

186. Freeman M, Baehler C, Spotts S. Automated laser-fluorescence sequencing. Biotechnology (N Y) 1990;8(2):147-148.

187. Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res 1987;15(11):4593-4602.

188. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 1987;238(4825):336-341.

189. Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 1970;227(259):680-685.

Anhang

Anhang A Suchsequenzen zur Initialisierung der LED

Protein	Organismus	Datenbank	Accesion Code	HFam
GUT esterase	Caenorhabditis briggsae	SWISS PROT	EST1_CAEBR	abH1.1
lipase	Rattus norvegicus	SWISS PROT	EST1_RAT	abH1.2
Bile-salt-activated lipase	Rattus norvegicus	SWISS PROT	BAL_RAT	abH1.3
acetylcholinesterase	Torpedo californica	SWISS PROT	ACES_TORCA	abH1.4
Para-nitrobenzyl esterase	Bacillus subtilis	SWISS PROT	PNBA_BACSU	abH1.5
Esterase B1	Culex pipiens	SWISS PROT	EST1_CULPI	abH1.6
Juvenile hormone esterase	Heliothis virescens	SWISS PROT	ESTJ_HELVI	abH1.7
Esterase 5B	Drosophila pseudoobscura	SWISS PROT	ESTB_DROPS	abH1.9
lipase 1	Yarrowia lipolytica	SWISS PROT	LIP1_YARLI	abH2.1
lipase 1	Candida rugosa	SWISS PROT	LIP1_CANRU	abH3.1
lipase 2	Moraxella sp.	SWISS PROT	LIP2_MORSP	abH4.1
esterase	Acinetobacter esterase	SWISS PROT	EST_ACICA	abH4.2
probable lipase/esterase	Pirellula sp. 1	GenBank	CAD77614	abH4.4
brefeldin A esterase	Bacillus subtilis	GenBank	AAC12774	abH6.1
lipase	Haemophilus influenzae	SWISS PROT	Y193_HAEIN	abH7.1
lipase 3	Moraxella sp.	SWISS PROT	LIP3_MORSP	abH7.2
epoxide hydrolase	Rattus norvegicus	SWISS PROT	HYES_RAT	abH8.3
epoxide hydrolase	Arabidopsis thaliana	PIR	T45731	abH8.4
2-hydroxy-6-phenyl-6-oxo-2,4-dienoic acid hydrolase	Pseudomonas azelaica	GenBank	AAB57641	abH8.8
Non-heme chloroperoxidase	Pseudomonas pyrrocinia	SWISS PROT	PRXC_PSEPY	abH8.13
CCG1-interacting factor B	Homo sapiens	SWISS PROT	C1IB_HUMAN	abH8.14
BioH protein	Escherichia coli	SWISS PROT	BIOH_ECOLI	abH9.2
Carboxylesterase	Geobacillus stearothermophilus	SWISS PROT	EST_BACST	abH11.1
lipase	Mycoplasma pneumoniae	SWISS PROT	ESL2_MYCPN	abH11.2
(S)-acetone-cyanohydrin lyase	Hevea brasiliensis	SWISS PROT	HNL_HEVBR	abH12.1
probable exported protein YPO1997	Yersinia pestis	PIR	AF0243	abH13.1
lysosomal acid lipase	Homo sapiens	SWISS PROT	LICH_HUMAN	abH14.1
lipase	Canis familiaris	SWISS PROT	LIPG_CANFA	abH14.2
lipase	Staphylococcus epidermidis	SWISS PROT	LIP_STAEP	abH15.1
lipase	Pseudomonas aeruginosa	SWISS PROT	LIP_PSEAE	abH15.2
lipase 2	Saccharomyces cerevisiae	SWISS PROT	TGL2_YEAST	abH15.3
triacylglycerol lipase	Propionibacterium acnes P-37	EMBL	CAA67627	abH16.1
probable lipase	Thermosynechococcus elongatus BP-1	GenBank	BAC09534	abH17.1
lipase	Bacillus subtilis	SWISS PROT	LIPA_BACSU	abH18.1
Palmitoyl-protein thioesterase 1	Bos taurus	SWISS PROT	PPT1_BOVIN	abH19.1
lipase	Rattus norvegicus	SWISS PROT	LIPH_RAT	abH20.1
lipoprotein lipase	Bos taurus	SWISS PROT	LIPL_BOVIN	abH20.2
pancreatic lipase	Sus scrofa	SWISS PROT	LIPP_PIG	abH20.3
serine esterase	Chlorobium tepidum TLS	GenBank	AAM73076	abH21.1
Carboxylesterase 2	Pseudomonas fluorescens	PDB	1AUR	abH22.1
lipase	Rhizomucor miehei	SWISS PROT	LIP_RHIMI	abH23.1
lipase	Saccharomyces cerevisiae	SWISS PROT	YJ77_YEAST	abH23.2
lipase	Pseudomonas fluorescens	SWISS PROT	LIPA_PSEFL	abH24.1
lipase 1	Moraxella sp.	SWISS PROT	LIP1_MORSP	abH25.1
cephalosporin-C deacetylase	Bacillus subtilis	PIR	G69596	abH26.1
Dipeptidyl peptidase IV	Homo sapiens	SWISS PROT	DPP4_HUMAN	abH27.1
Prolyl endopeptidase	Sus scrofa	SWISS PROT	PPCE_PIG	abH28.1
X-Pro dipeptidyl-peptidase	Lactococcus lactis	SWISS PROT	PEPX_LACLC	abH29.1
Cocaine esterase	Rhodococcus sp. MB1 'Bresler 1999'	SWISS PROT	COCE_RHOSM	abH30.1
carboxymethylenebutenolidase	Pseudomonas aeruginosa	PIR	T44669	abH31.2
Endo-1,4-beta-xylanase Z	Clostridium thermocellum	SWISS PROT	XYNZ_CLOTM	abH32.1
Endo-1,4-beta-xylanase Y	Clostridium thermocellum	SWISS PROT	XYNY_CLOTM	abH32.2
Antigen 85-C	Mycobacterium tuberculosis	SWISS PROT	A85C_MYCTU	abH33.1
Antigen 85-B	Mycobacterium bovis	SWISS PROT	A85B_MYCBO	abH33.3
Lysosomal protective protein	Homo sapiens	SWISS PROT	PRTP_HUMAN	abH34.1
Carboxypeptidase D	Triticum aestivum	SWISS PROT	P08819_1	abH34.2
Acyl transferase	Vibrio harveyi	SWISS PROT	LUXD_VIBHA	abH35.1
cutinase	Colletotrichum glosporioides	SWISS PROT	CUTI_COLGL	abH36.1
cutinase	Fusarium solani	SWISS PROT	CUTI_FUSSO	abH36.3
cutinase	Mycobacterium tuberculosis	SWISS PROT	CUT1_MYCTU	abH36.4
lipase B	Candida antarctica	SWISS PROT	LIPB_CANAR	abH37.1

Anhang B Streng konservierte Positionen (SKPs)

Tabelle 15: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH9, Positionen der SKPs in der Referenzstruktur 1QO7, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	1QO7	Sekundärstruktur	abH9	Konservierung	LMZ rel.
1	P110	E	L(16%), P(78%)	3,43	7,9
2	I111	E	A(13%), I(15%), L(25%), M(6%), V(39%)	1,71	0
3	A112	E	I(22%), L(42%), V(25%)	2,14	0,2
4	L114	E	H(12%), I(10%), L(42%), T(7%), V(13%)	0,45	0,4
5	H115		G(16%), H(84%)	5,20	1,4
6	G116		G(84%), W(16%)	3,94	5,3
7	W117	S	G(55%), W(39%)	2,50	0,4
8	F142	E	F(49%), W(9%), Y(33%)	3,33	4,1
9	H143	E	D(9%), E(19%), H(16%), Q(6%), R(33%)	0,76	6,1
10	L144	E	I(25%), L(24%), V(49%)	2,72	0,1
11	V145	E	H(10%), I(37%), L(9%), V(36%)	1,53	0
12	V146	E	A(7%), C(6%), I(10%), L(51%), V(18%)	1,60	0
13	P147	E	F(34%), L(6%), P(36%), V(15%)	0,08	2
14	S148		D(60%), S(36%)	2,64	0,4
15	L149		I(10%), L(36%), M(7%), Q(45%)	0,88	1
16	P150	T	P(52%), R(45%)	1,88	0,6
17	G151	T	G(94%)	5,64	0
18	A170	H	A(37%), I(6%), V(37%)	0,81	0
19	V172	H	A(6%), D(45%), I(13%), M(9%), V(19%)	0,13	1,5
20	V173	H	A(6%), F(15%), I(28%), L(7%), M(24%), V(9%), W(6%)	0,97	2
21	Q175	H	A(16%), E(9%), K(30%), N(13%), R(18%)	0,71	40,1
22	L176	H	I(15%), L(75%), V(10%)	3,06	0,4
23	M177	H	A(9%), M(37%), R(45%)	1,07	1,8
24	D179	H	A(9%), H(25%), M(7%), R(33%)	0,58	0
25	G185		K(33%), Q(6%), R(24%)	1,90	1,8
26	Y186	E	A(10%), F(15%), W(43%), Y(21%)	2,51	0
27	I188	E	A(10%), I(15%), L(22%), V(42%), W(7%)	1,29	0
28	G190	E	A(6%), G(90%)	5,29	0
29	G191		G(79%), W(18%)	3,37	14,3
30	D192	T	D(37%), S(63%)	2,41	13,1
31	I193	H	L(18%), W(72%)	5,18	3,1
32	G194	H	G(100%)	6,00	0
33	S195	H	A(12%), G(18%), S(67%)	2,21	5
34	F196	H	A(6%), I(10%), L(22%), M(6%), T(36%), V(6%)	0,40	31,5
35	V197	H	I(19%), L(49%), V(19%)	1,96	0
36	R199	H	L(52%), R(9%), S(18%), T(12%)	0,48	6,9
37	L201	H	A(7%), I(9%), L(24%), M(13%), Y(45%)	1,40	6,9
38	C208	E	C(16%), V(73%)	2,26	0
39	V211	E	I(7%), L(55%), M(6%), V(13%), Y(7%)	1,51	0
40	H212	E	H(37%), I(18%), V(33%)	0,70	0
41	L213	E	L(64%), M(7%), S(7%), T(13%)	1,66	0,9
42	N214	S	N(33%), R(36%), T(7%), V(12%)	0,50	7,1
43	L215		A(19%), F(9%), G(40%), M(19%)	0,27	6,8
44	C216		A(9%), C(15%), I(33%), L(12%), S(13%), V(9%)	0,31	4,1
45	F342	E	C(9%), F(9%), I(36%), L(16%), R(7%), V(6%)	0,16	3,6
46	S343	E	A(24%), I(18%), L(10%), S(12%), V(33%)	0,97	0
47	F345		F(30%), G(63%)	1,59	1,6
48	D348	S	D(72%), E(28%)	4,29	7,6
49	I357	H	A(15%), L(67%), V(10%)	1,94	0,1
50	T360	G	A(40%), K(21%), L(10%), T(9%)	0,23	0
51	H374	S	H(99%)	7,82	6
52	F375		A(25%), F(37%), S(22%)	0,39	1,2
53	L388	H	A(34%), F(7%), I(22%), L(19%), V(13%)	0,84	1,6
54	V392	H	A(9%), F(24%), I(12%), L(7%), M(12%), V(25%)	0,65	1,1

Tabelle 16: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH22, Positionen der SKPs in der Referenzstruktur 1FJ2, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	1FJ2	Sekundärstruktur	abH22	Konservierung	LMZ rel.
1	A18	E	A(12%), C(12%), F(12%), S(9%), T(39%), V(6%)	0,59	0
2	V19	E	I(24%), L(24%), V(52%)	2,77	0,1
3	I20	E	I(67%), L(12%), V(18%)	3,02	0
4	L22	E	I(6%), L(85%)	3,16	0
5	H23		H(100%)	8,00	2,3
6	G24		G(100%)	6,00	0
7	L25	S	L(88%), S(12%)	2,72	21
8	I44	E	F(6%), I(42%), M(6%), T(21%), V(18%)	1,40	0,1
9	K45	E	K(73%), R(21%)	3,38	22,7
10	Y46	E	F(33%), I(12%), V(9%), W(30%), Y(15%)	2,17	0,5
11	I47	E	A(6%), I(67%), L(9%), V(12%)	2,40	1,8
12	C48	E	C(45%), F(24%), L(21%), Y(9%)	1,75	0,5
13	P49		P(100%)	7,00	7,2
14	H50		H(33%), N(6%), Q(12%), S(9%), T(39%)	1,24	26,1
15	A51		A(97%)	3,82	2
16	P52		N(6%), P(88%), S(6%)	5,13	31
17	V53	E	E(9%), N(6%), P(9%), R(9%), S(30%), T(24%), V(6%)	0,47	64
18	A88	H	A(45%), C(6%), Q(6%), S(9%), T(6%), V(15%)	0,64	2
19	N90	H	E(6%), H(18%), N(18%), Q(6%), R(6%), S(15%), T(6%), V(6%), Y(15%)	0,31	61,7
20	I91	H	F(6%), I(24%), L(18%), V(52%)	2,50	1,8
21	A93	H	A(30%), D(9%), E(18%), G(6%), K(9%), N(15%), Q(6%)	0,38	46,2
22	L94	H	F(6%), I(12%), L(67%), M(12%)	2,55	13,3
23	I95	H	I(70%), L(18%), V(9%)	2,98	0
24	Q97	H	A(18%), E(18%), H(9%), K(15%), Q(12%), S(6%), T(12%)	0,35	50,3
25	R107	G	K(12%), N(18%), R(61%)	2,54	15,7
26	I108	E	I(73%), L(9%), V(15%)	3,40	0
27	L110	E	I(33%), L(39%), V(24%)	2,45	0
28	G112	E	G(100%)	6,00	0
29	F113	E	F(94%), L(6%)	5,31	3
30	S114	T	S(100%)	4,00	6,6
31	Q115	H	M(27%), Q(73%)	3,02	1,6
32	G116	H	G(100%)	6,00	0
33	G117	H	A(21%), G(76%)	3,49	0
34	A118	H	A(76%), C(6%), S(6%), V(9%)	2,39	0
35	L119	H	I(9%), L(45%), M(15%), T(9%), V(21%)	1,91	0
36	L121	H	I(12%), L(76%), M(6%)	2,90	0,3
37	T123	H	A(12%), L(6%), N(6%), S(39%), T(36%)	1,60	0
38	L131		I(9%), L(76%), V(6%), Y(6%)	2,60	3,6
39	V134	E	C(6%), G(9%), I(27%), V(48%)	1,40	0,2
40	T135	E	A(15%), F(9%), I(12%), L(9%), M(9%), T(6%), V(39%)	1,17	0
41	A136	E	A(55%), G(27%), V(18%)	1,47	0
42	L137	E	F(21%), H(6%), L(70%)	2,05	0,1
43	S138	S	S(97%)	3,83	0
44	C139		C(30%), G(42%), S(12%), T(12%)	0,96	0
45	Q163	E	H(9%), L(21%), M(9%), Q(33%), W(15%)	0,09	0
46	C164	E	A(9%), C(42%), G(9%), L(12%), S(9%)	0,92	0
47	G166	E	G(94%)	5,45	0
48	D169		D(97%)	5,59	9,9
49	G177	H	A(6%), G(91%)	4,98	0
50	T180	H	A(24%), G(9%), S(30%), T(36%)	1,55	0,1
51	H203	S	H(97%)	8,00	23
52	S204	S	E(18%), S(58%), Y(12%)	1,32	39,8
53	V213	H	I(18%), L(30%), M(6%), V(36%)	2,08	0
54	I217	H	I(33%), L(58%)	2,75	0,6

Tabelle 17: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH33, Positionen der SKPs in der Referenzstruktur 1F0N, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	1F0N	Sekundästruktur	abH33	Konservierung	LMZ rel.
1	A35	E	A(65%), E(5%), G(5%), M(10%), T(10%), V(5%)	1,51	0,8
2	V36	E	I(15%), L(40%), V(45%)	2,55	0
3	Y37	E	W(5%), Y(95%)	6,54	1,5
4	L39	E	L(100%)	4,00	0
5	D40		D(85%), G(5%), N(10%)	4,50	4,3
6	G41		G(100%)	6,00	9,4
7	L42	T	A(5%), G(10%), I(5%), L(55%), M(20%), V(5%)	1,45	16,9
8	L66		A(5%), I(30%), L(45%), V(20%)	2,35	0,3
9	S67	E	I(5%), N(15%), S(65%), T(15%)	2,15	0
10	I68	E	I(15%), L(10%), V(75%)	3,27	0,3
11	V69	E	A(10%), I(35%), V(55%)	2,83	0
12	M70	E	I(10%), L(5%), M(85%)	4,02	0
13	P71	E	P(100%)	7,00	0,1
14	V72		I(5%), L(5%), M(10%), Q(5%), T(5%), V(70%)	2,33	3,6
15	G73		A(5%), E(10%), G(80%), L(5%)	3,23	6,3
16	G74		A(10%), G(90%)	4,90	3,9
17	Q75	T	A(15%), E(5%), G(5%), K(5%), Q(65%), S(5%)	2,09	24,9
18	T94		A(5%), I(5%), K(5%), N(5%), P(5%), R(10%), T(65%)	1,78	35,8
19	K96	B	K(80%), M(10%), Q(10%)	3,30	29,9
20	W97	H	W(100%)	11,00	0,2
21	T99	H	T(100%)	5,00	32,9
22	F100	H	F(100%)	6,00	0
23	L101	H	I(5%), L(95%)	3,81	0
24	S103	T	E(5%), H(5%), K(15%), Q(20%), R(5%), S(50%)	1,41	44,7
25	G119	S	D(5%), G(65%), H(5%), K(10%), N(5%), Q(10%)	1,98	10,3
26	S120	E	N(20%), R(25%), S(55%)	1,71	1,1
27	A122	E	A(45%), I(35%), V(20%)	1,57	0
28	G124	E	G(100%)	6,00	0
29	L125	E	I(15%), L(70%), M(15%)	3,05	1,5
30	S126	T	S(100%)	4,00	14,3
31	M127	H	M(100%)	5,00	5
32	A128	H	A(45%), G(20%), S(35%)	1,86	0
33	G129	H	A(20%), G(80%)	4,00	0
34	S130	H	G(15%), L(5%), S(55%), T(25%)	1,60	5
35	S131	H	A(15%), G(5%), S(75%), T(5%)	2,66	6,2
36	M133	H	L(70%), M(15%), V(15%)	2,84	0
37	L135	H	F(10%), I(5%), L(85%)	3,13	1,1
38	F143	E	F(85%), Y(15%)	5,26	0,1
39	A146	E	A(65%), S(5%), T(5%), V(25%)	1,98	0
40	G147	E	A(20%), F(5%), G(65%), I(5%), M(5%)	2,01	0
41	S148	E	A(5%), G(10%), S(85%)	3,05	0
42	L149	E	F(15%), I(5%), L(55%), M(10%), V(5%), W(5%), Y(5%)	1,71	0
43	S150	S	A(5%), S(95%)	3,72	0,7
44	A151		A(40%), G(60%)	2,80	0,8
45	V209	E	A(10%), I(15%), M(5%), V(70%)	2,78	0
46	Y210	E	A(15%), F(5%), S(15%), Y(65%)	2,55	1,7
47	G212		A(5%), G(95%)	5,43	0
48	E230	H	E(95%), L(5%)	4,24	0
49	N237	H	N(65%), S(10%), T(25%)	3,07	0
50	F240	H	F(70%), I(5%), L(25%)	3,25	0,1
51	H262	S	H(100%)	8,00	17,7
52	S263	S	A(5%), D(10%), G(5%), N(5%), S(75%)	2,47	15,6
53	M274	H	A(10%), F(5%), G(5%), M(65%), S(10%), T(5%)	1,67	0
54	L278	H	F(10%), I(20%), L(50%), M(10%), V(10%)	2,17	0

Tabelle 18: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH27, Positionen der SKPs in der Referenzstruktur 1J2E, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	1J2E	Sekundärstruktur	abH27	Konservierung	LMZ rel.
1	L542	E	L(39%), M(7%), T(15%), V(33%)	1,57	0
2	L543	E	F(7%), I(15%), L(57%), V(17%)	2,22	0
3	L544	E	F(20%), I(15%), L(30%), M(7%), V(26%)	1,56	0
4	V546	E	I(9%), P(7%), Q(13%), V(57%)	1,13	2,4
5	Y547		D(11%), Y(85%)	4,42	36,8
6	G549		G(85%)	3,97	0,7
7	P550	T	P(100%)	7,00	2,5
8	I573		A(9%), F(7%), I(20%), V(13%), Y(48%)	1,73	2,1
9	I574	E	A(7%), I(52%), L(7%), V(28%)	2,47	0
10	V575	E	I(11%), T(7%), V(70%), Y(9%)	2,40	2,9
11	A576	E	A(33%), F(13%), I(9%), L(9%), V(24%), W(7%)	0,43	0,2
12	S577	E	C(13%), K(13%), Q(9%), R(13%), S(22%), T(15%)	0,06	3
13	F578	E	F(15%), I(28%), L(22%), V(30%)	1,74	1,4
14	D579		D(98%)	5,83	4
15	G580		G(61%), N(24%)	2,14	5,7
16	R581	T	R(98%)	4,66	32,4
17	G582	T	G(98%)	5,70	0,1
18	V603	H	A(17%), I(9%), L(7%), S(9%), V(50%)	1,11	2,5
19	D605	H	D(98%)	5,74	2,8
20	Q606	H	Q(93%)	4,13	1
21	I607	H	I(43%), L(28%), V(17%)	2,14	1,2
22	E608	H	A(17%), E(30%), Q(7%), T(15%)	0,10	26,9
23	A609	H	A(52%), G(17%), V(22%)	1,14	0,9
24	R611	H	A(7%), E(7%), K(33%), Q(9%), R(30%)	1,33	42,2
25	R623	E	H(13%), K(11%), R(76%)	3,39	18,3
26	I624	E	I(54%), L(9%), M(7%), V(28%)	2,82	1,9
27	I626	E	I(63%), V(30%)	3,13	0
28	G628	E	G(100%)	6,00	1,5
29	W629	E	K(9%), W(87%)	7,71	21,4
30	S630	T	G(11%), S(85%)	2,95	17,6
31	Y631	H	F(7%), N(9%), Y(85%)	5,11	2,3
32	G632	H	G(98%)	5,66	0
33	G633	H	G(98%)	5,57	0
34	Y634	H	A(7%), I(7%), L(28%), M(26%), T(9%), Y(15%)	0,73	0
35	V635	H	A(20%), S(22%), T(41%), V(15%)	1,40	0
36	S637	H	K(26%), L(7%), M(35%), S(20%)	0,53	0
37	V639	H	L(61%), M(7%), V(22%)	2,23	1,7
38	F647		F(76%), Y(13%)	4,12	4,2
39	G650	E	A(20%), G(72%)	3,08	0
40	I651	E	A(13%), I(43%), M(15%), S(7%), V(20%)	1,44	0
41	A652	E	A(59%), S(26%), V(7%)	1,81	0
42	V653	E	G(22%), I(9%), L(7%), V(57%)	0,90	0
43	A654	S	A(83%), S(11%)	2,85	0
44	P655		P(98%)	6,66	0
45	L702	E	I(22%), L(65%), V(11%)	2,86	4
46	I703	E	I(52%), L(7%), M(13%), V(17%)	1,96	0,2
47	G705	E	G(93%)	5,25	0,1
48	D708		D(100%)	6,00	0,6
49	S716	H	A(15%), S(54%), T(28%)	2,12	3,3
50	I719	H	F(11%), I(17%), L(65%)	2,33	1,3
51	H740	T	H(98%)	7,74	12,6
52	G741	T	G(35%), S(41%)	1,07	48,8
53	M755	H	A(7%), I(13%), L(35%), M(35%)	1,75	0
54	I759	H	F(15%), I(13%), L(54%)	1,58	0

Tabelle 19: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH4, Positionen der SKPs in der Referenzstruktur 1LZL, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	1LZL	Sekundästruktur	abH4	Konservierung	LMZ rel.
1	V81	E	A(22%), C(6%), G(12%), L(21%), T(6%), V(24%)	0,24	0
2	L82	E	F(6%), I(30%), L(29%), V(25%)	1,88	0
3	L83	E	F(6%), I(18%), L(29%), M(6%), V(39%)	1,99	0
4	I85	E	F(31%), I(22%), L(12%), V(6%), Y(22%)	1,54	0
5	H86		H(95%)	7,27	0,9
6	G87		G(95%)	5,30	9,5
7	G88	S	G(95%)	5,37	16,2
8	F111		A(32%), C(16%), F(10%), S(11%), T(5%), V(10%), Y(5%)	0,24	8,5
9	A112	E	A(14%), I(9%), L(8%), P(9%), Q(6%), T(12%), V(32%)	0,25	0,9
10	V113	E	A(6%), I(10%), V(74%)	2,60	0
11	A114	E	A(6%), F(6%), I(21%), L(20%), V(40%)	1,72	0
12	N115	E	A(10%), F(6%), N(6%), S(57%)	1,06	0
13	V116	E	A(5%), I(21%), V(66%)	2,74	0
14	E117		D(61%), E(8%), G(10%), N(12%)	2,50	39,9
15	Y118		Y(98%)	6,68	2,8
16	R119		R(80%), S(6%), T(9%)	2,86	28,2
17	L120		K(10%), L(78%), P(5%)	1,78	22,8
18	V130	H	A(5%), F(6%), H(5%), I(21%), L(29%), V(22%), Y(5%)	0,89	0,3
19	D132	H	D(79%), E(14%), Q(5%)	4,29	0,5
20	C133	H	A(30%), C(42%), G(5%), V(15%)	1,73	0
21	A135	H	A(41%), D(19%), N(8%), S(6%), T(8%)	0,24	2,1
22	A136	H	A(75%), V(15%)	2,30	0
23	L137	H	A(6%), F(5%), I(8%), L(30%), T(20%), V(11%), W(5%), Y(9%)	0,26	0
24	Y139	H	A(5%), F(5%), H(8%), W(64%), Y(11%)	4,30	12,8
25	R153	E	K(11%), R(74%)	2,95	19,3
26	I154	E	I(46%), L(20%), V(25%)	2,41	0
27	V156	E	F(5%), I(16%), L(24%), V(49%)	2,09	0
28	G158	E	G(100%)	6,00	0
29	Q159	E	D(59%), E(14%), N(5%)	1,98	30
30	S160	T	S(100%)	4,00	6,2
31	A161	H	A(74%), V(12%)	2,16	0
32	G162	H	G(100%)	6,00	0
33	G163	H	A(15%), G(85%)	4,43	0
34	G164	H	G(12%), H(8%), N(56%), T(6%)	1,50	1,7
35	L165	H	A(5%), I(10%), L(66%), M(8%)	2,15	0,4
36	A167	H	A(56%), F(5%), L(9%), T(18%)	1,01	0,3
37	T169	H	A(5%), I(10%), L(20%), T(19%), V(41%)	1,36	0
38	V181		I(28%), L(21%), P(21%), V(11%)	0,11	16,5
39	Q184	E	I(5%), Q(71%), V(5%)	1,83	0,5
40	F185	E	F(5%), I(21%), L(20%), M(9%), V(30%)	1,20	0,6
41	L186	E	A(5%), L(76%), P(5%)	1,89	0
42	E187	E	A(8%), F(5%), I(44%), L(14%), V(9%), Y(5%)	0,64	18,8
43	I188	S	W(9%), Y(66%)	2,97	3
44	P189		A(5%), P(86%)	4,83	0,4
45	L254	E	I(41%), L(24%), V(29%)	2,48	0
46	S255	E	A(18%), G(5%), I(22%), L(11%), Q(5%), V(25%)	0,44	3,9
47	M257	E	A(57%), C(18%), G(14%)	1,39	8,7
48	D260		D(100%)	6,00	0,4
49	G266	H	A(9%), G(57%), S(12%)	1,64	0
50	Y269	H	F(10%), L(12%), M(5%), Y(70%)	3,64	1,3
51	H290	T	H(98%)	7,53	8,8
52	G291	T	A(11%), D(11%), G(68%)	2,28	12,1
53	A308	H	A(18%), G(8%), I(24%), L(6%), M(5%), T(9%), V(16%)	0,00	7,3
54	I312	H	F(14%), I(21%), L(49%), V(9%)	1,76	0

Tabelle 20: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH1, Positionen der SKPs in der Referenzstruktur 1AKN, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	1AKN	Sekundärstruktur	abH1	Konservierung	LMZ rel.
1	V100	E	I(5%), V(90%)	3,64	0
2	M101	E	I(9%), L(23%), M(58%), V(7%)	2,78	0,1
3	I102	E	I(11%), V(77%)	2,94	0
4	I104	E	F(10%), I(76%), L(5%), V(9%)	2,92	0,3
5	Y105		F(5%), H(62%), Y(30%)	4,36	0,5
6	G106		G(96%)	5,50	0
7	G107	S	G(96%)	5,49	8,5
8	V136		I(12%), L(7%), V(71%)	2,71	0
9	I137	E	I(42%), V(54%)	3,38	0
10	V138	E	F(6%), L(23%), V(60%)	1,96	0
11	V139	E	A(9%), I(8%), V(83%)	3,19	0,1
12	T140	E	A(7%), N(7%), S(28%), T(47%)	1,60	0,4
13	F141	E	F(15%), I(39%), L(10%), M(14%), P(7%), V(10%)	1,14	0,5
14	N142		A(5%), N(51%), Q(32%), S(6%)	1,95	0,7
15	Y143		Y(93%)	6,36	0
16	R144		R(99%)	5,00	5,3
17	V145		L(69%), V(25%)	2,63	1,1
18	L164	H	F(6%), H(8%), L(69%), M(5%), Y(7%)	1,84	0
19	D166	H	D(100%)	6,00	0
20	Q167	H	I(7%), M(6%), Q(80%)	2,74	0
21	M169	H	A(28%), E(5%), L(40%), M(15%)	0,65	15,6
22	A170	H	A(93%)	3,45	0
23	I171	H	I(10%), L(82%)	3,23	0
24	W173	H	F(8%), W(92%)	9,56	2,2
25	N187	E	K(10%), N(30%), Q(7%), R(21%), S(17%)	0,61	11
26	I188	E	I(34%), V(61%)	3,22	0
27	L190	E	I(36%), L(42%), V(18%)	2,53	0
28	G192	E	G(98%)	5,82	0
29	E193	E	E(56%), H(16%), Q(16%)	2,02	0
30	S194	T	S(94%)	3,51	10,4
31	A195	H	A(87%), S(6%)	3,03	0,4
32	G196	H	G(100%)	6,00	0
33	G197	H	A(40%), G(47%), S(11%)	2,11	0
34	A198	H	A(40%), G(7%), M(7%), S(8%), T(12%), V(5%)	0,22	0
35	S199	H	A(6%), L(5%), S(78%)	2,11	0
36	S201	H	A(7%), D(6%), G(8%), H(20%), N(6%), S(37%)	0,48	8,3
37	Q203	H	H(36%), L(41%), M(8%), Q(11%)	0,73	0
38	I213		F(81%), I(6%), V(6%)	3,80	6
39	A216	E	A(68%), G(10%), V(7%)	1,63	0
40	I217	E	I(86%), V(8%)	3,53	0
41	S218	E	A(7%), I(7%), L(24%), M(20%), P(6%), S(20%), V(10%)	0,34	1,3
42	Q219	E	E(19%), M(28%), Q(37%)	1,10	0
43	S220	S	S(92%)	3,50	0,3
44	G221		A(5%), G(85%)	4,17	2,2
45	A314	E	A(6%), I(34%), L(14%), M(6%), T(7%), V(23%)	1,19	0,4
46	G315	E	G(84%)	3,86	1,6
47	N317	E	N(45%), T(24%), V(14%)	0,97	0,1
48	D320	T	D(6%), E(83%)	3,34	0
49	T397	H	A(21%), I(10%), L(10%), S(13%), T(20%), V(11%)	0,20	3,9
50	A400	H	A(8%), F(17%), L(12%), M(7%), T(14%), V(22%), Y(7%)	0,07	11,5
51	H435	T	H(88%)	5,91	3,2
52	A436	T	A(26%), C(6%), G(45%), S(6%)	0,94	1,4
53	W468	H	F(14%), I(6%), L(7%), V(7%), W(55%)	2,68	1
54	A472	H	A(72%), I(6%), S(6%), V(8%)	2,10	0

Tabelle 21: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH15, Positionen der SKPs in der Referenzstruktur 4LIP, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	4LIP	Sekundärstruktur	abH15	Konservierung	LMZ rel.
1	P10		P(100%)	7,00	3,6
2	I11	E	I(70%), L(13%), V(13%)	3,07	1,7
3	I12	E	I(30%), L(13%), V(57%)	3,05	0
4	V14	E	A(10%), E(7%), S(7%), V(70%)	1,69	0
5	H15		H(90%), P(10%)	6,19	0
6	G16		G(100%)	6,00	0
7	L17	T	F(30%), I(13%), L(37%), M(17%)	1,82	37,5
8	A42		A(60%), E(7%), Y(27%)	1,10	7,5
9	T43		E(10%), K(10%), N(20%), Q(13%), R(17%), S(7%), T(20%)	0,82	43,2
10	V44	E	T(10%), V(87%)	3,06	4,7
11	Y45	E	F(23%), H(10%), L(7%), S(7%), Y(47%)	2,44	30,6
12	V46	E	A(10%), E(20%), L(7%), T(10%), V(50%)	0,85	17,1
13	A47		A(57%), P(13%), S(10%), T(13%)	1,36	0
14	N48		D(7%), E(7%), N(17%), Q(17%), S(30%), T(10%)	0,73	34,4
15	L49		I(7%), L(27%), Q(13%), S(7%), V(47%)	1,13	2,2
16	S50		A(13%), G(23%), S(60%)	1,93	0
17	G51	S	A(50%), G(10%), P(13%), Q(7%), S(17%)	1,13	2,9
18	G62	H	A(37%), G(63%)	2,94	0
19	Q64	H	E(40%), Q(57%)	3,25	24,3
20	L65	H	L(97%)	3,68	0
21	A67	H	A(47%), L(7%), Y(23%)	0,42	52,3
22	Y68	H	Q(40%), Y(50%)	2,12	26,8
23	V69	H	I(33%), L(13%), V(47%)	2,57	0
24	T71	H	A(10%), E(13%), G(33%), Q(10%), S(7%), T(10%)	0,27	57,3
25	K80		K(77%), P(13%), R(10%)	3,16	23
26	V81	E	I(27%), V(73%)	3,61	0
27	L83	E	F(10%), L(87%)	3,18	0
28	G85	E	A(7%), G(93%)	5,24	0
29	H86	E	H(93%), Y(7%)	7,25	0,9
30	S87	T	S(100%)	4,00	7,7
31	Q88	H	A(7%), H(20%), M(23%), Q(50%)	1,52	0,1
32	G89	H	G(100%)	6,00	0
33	G90	H	A(13%), G(63%), S(20%)	2,59	0
34	L91	H	L(33%), P(20%), Q(43%)	0,54	1,8
35	T92	H	S(7%), T(83%), V(7%)	3,61	0
36	R94	H	R(100%)	5,00	3,1
37	V96	H	A(10%), L(33%), V(53%)	1,99	0
38	V104	E	I(13%), M(7%), N(17%), S(7%), V(50%)	0,78	0,3
39	V107	E	E(23%), G(7%), V(60%)	0,74	0
40	T108	E	I(17%), T(63%)	1,82	0
41	T109	E	A(10%), D(7%), E(20%), S(30%), T(30%)	1,22	0,3
42	I110	E	I(30%), M(7%), V(30%), Y(23%)	1,23	0
43	G111	S	A(10%), G(50%), N(7%), Q(20%), S(7%), T(7%)	1,21	0,6
44	T112		A(7%), G(33%), K(17%), N(7%), T(23%)	0,36	0,2
45	S208	E	S(83%), T(13%)	3,08	0,4
46	W209	E	F(7%), W(60%), Y(27%)	5,09	1,2
47	G211	E	G(80%), T(13%)	3,52	1,6
48	D264	S	D(100%)	6,00	0,7
49	A272	H	A(17%), M(13%), S(57%), T(7%)	1,55	1,3
50	G275	S	G(67%), P(33%)	2,56	1,2
51	H286	T	H(100%)	8,00	9,9
52	L287	T	F(10%), I(10%), L(57%), V(13%)	1,50	23,8
53	H311	H	C(7%), H(57%), I(7%), L(13%), Q(10%)	1,65	1,8
54	L315	H	L(83%), V(10%)	3,05	0,4

Tabelle 22: Prozentualer Anteil der Aminosäuren (größer 5%) an den SKPs der Familie abH14, Positionen der SKPs in der Referenzstruktur 1K8Q, sowie deren Konservierung C und relative Lösungsmittelzugänglichkeit (LMZ rel.).

SKP	1K8Q	Sekundästruktur	abH14	Konservierung	LMZ rel.
1	V60	E	P(25%), V(65%)	1,52	0,7
2	A61	E	A(8%), I(12%), V(79%)	3,06	0
3	F62	E	F(38%), L(40%), Y(21%)	2,17	0,1
4	Q64	E	H(6%), L(10%), M(13%), Q(60%), V(10%)	1,46	3,7
5	H65		H(96%)	7,29	1,2
6	G66		G(96%)	5,55	3,9
7	L67	T	L(87%)	3,29	33,9
8	Y92		F(29%), Y(71%)	5,28	5,9
9	D93	E	D(98%)	5,85	4,2
10	V94	E	V(96%)	3,93	0
11	W95	E	W(96%)	10,25	0
12	L96	E	L(63%), M(33%)	3,08	0
13	G97		A(6%), G(88%)	4,44	0,6
14	N98		N(96%)	5,63	2,5
15	S99		A(13%), F(8%), M(6%), N(17%), S(37%), V(10%)	0,33	5,6
16	R100	T	R(100%)	5,00	0,8
17	G101	T	G(96%)	5,34	0
18	A127	H	A(48%), G(31%), I(10%), S(6%)	1,14	0,6
19	Y129	T	E(6%), F(12%), Y(73%)	4,14	20,5
20	L131	H	I(17%), L(79%)	3,26	0,6
21	A133	H	A(79%), S(8%)	2,36	18
22	T134	H	I(10%), L(6%), M(33%), S(10%), T(27%), V(6%)	0,76	3,1
23	I135	H	I(81%), V(13%)	3,38	0
24	F137	H	F(29%), H(8%), K(10%), L(8%), Y(42%)	2,04	27,4
25	K146		A(6%), K(38%), Q(33%), S(15%)	1,52	40,2
26	L147	E	I(29%), L(56%), V(13%)	2,74	1
27	Y149	E	F(6%), L(10%), Y(77%)	4,08	0
28	G151	E	G(90%)	4,71	0
29	H152	E	F(10%), H(81%), Y(8%)	5,36	4,6
30	S153	T	S(96%)	3,71	9,4
31	Q154	H	L(10%), Q(88%)	3,67	3,4
32	G155	H	G(98%)	5,77	0
33	T156	H	C(10%), G(6%), T(73%)	2,38	0
34	T157	H	A(13%), L(19%), T(54%)	1,33	0,1
35	I158	H	A(12%), I(35%), M(6%), Q(12%), S(10%), T(12%), V(12%)	0,42	0,7
36	F160	H	F(67%), L(19%), W(6%)	3,16	0,3
37	A162	H	A(38%), H(6%), L(15%), M(17%), R(6%)	0,20	1,7
38	I173	E	I(73%), L(6%), V(15%)	3,35	4,3
39	F176	E	A(13%), F(38%), L(6%), M(15%), Y(12%)	0,85	0
40	Y177	E	F(25%), H(13%), I(19%), Q(10%), Y(15%)	0,41	0,1
41	A178	E	A(62%), G(6%), L(19%), M(6%)	1,45	0
42	L179	E	I(6%), L(83%), M(6%)	3,41	0
43	A180	S	A(79%), G(8%), S(8%)	2,88	0
44	P181		P(96%)	7,00	1,1
45	V318	E	I(8%), L(56%), M(19%), V(17%)	2,54	0
46	W319	E	F(19%), V(6%), W(29%), Y(37%)	2,66	0
47	G321	E	G(62%), S(31%)	2,57	0
48	D324		D(100%)	6,00	0,9
49	V332	H	I(21%), L(6%), V(69%)	3,15	0,4
50	L335	H	F(6%), H(8%), I(8%), L(62%), M(8%)	1,51	1,7
51	H353	T	H(100%)	8,00	7
52	L354	T	F(13%), I(10%), L(56%), M(12%)	1,72	2,4
53	I370	H	I(50%), L(12%), M(12%), V(19%)	2,10	0
54	M374	H	C(8%), F(6%), I(21%), L(19%), M(37%), V(8%)	1,66	3

Anhang C Konservierte Sequenzblöcke des Sequenzdatensatzes S_1 für das Design der familienspezifischen Primer bestimmt mit MOTIF.[167]

	Block $1M_1$		Block $1M_2$		Block $1M_3$	
01aPCD_ARTOX	7	AHTTAGDLGG	25	FRGVPYAEPPVGDLRWRAARPHAGW	84	EDCLTLNLWTPNLDGGSRPVLV
02aPNBA_BACSU	6	VTTQYGKVKG	24	WKGIPYAKPPVGQWRFKAPEPPEVW	80	EDCLYVNVFAPDTPSQNLPVMV
04aEST1_CAEBR	20	VNTNYGKVEG	38	FLAIPFAKPPVDNLRFEKPEAPEPW	91	EDCLTLNVIKPKTIEKKLPVLF
04bEST1_CAEEL	20	VETNYGKVEG	38	FLAIPFAKPPVDDLRFEKPVAPDPW	91	EDCLTLNIIKPKKAEKLPVLFW
06aCRYS_DICDI	36	VLLSDGAIRG	54	FYGIPFARPPIDELRYEDPQPPKPW	109	EDCLYLDVFIPRTVNPGSKVPV
06bD2_DICDI	34	VVTQFGAIKG	52	FYGIPFAQPPVNQLRWENPIDLKPW	107	EDCLYLDVFTPKDATPNSKYPV
09aEST1_RAT	26	VDTTKGKVLG	50	FLGVPFAKPPLGSLRFAPPEPAEPW	114	EDCLYLNIYSPADLTKNSRLPV
09bEST1_MESAU	34	RNTHTGQVRG	58	FLGIPFAKPPVGPLRFAPPEPPEPW	120	EDCLYLNIYTPAHAHEGSNLPV
09cEST2_RABIT	8	RNTHTGQVRG	32	FLGIPFAKPPLGPLRFAPPEPAEAW	94	EDCLYLNIYSPAHAREGSDLPV
09dESTN_MOUSE	26	VDTTQGKVLG	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYSPADLTKSSQLPV
09eEST1_HUMAN	26	VDTVHGKVLG	50	FLGIPFAKPPLGPLRFTPPQPAEPW	114	EDCLYLNIYTPADLTKKNRLPV
09fEST1_RABIT	8	VDTVKGKVLG	32	FLGVPFAKPPLGSLRFAPPQPAESW	96	EDCLYLNIYTPADLTKRGRLPV
09gES10_RAT	26	VNTVKGKVLG	50	FLGIPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNVYTPADLTKNSRLPV
09hES22_MOUSE	27	VDTVQGKVLG	51	FLGVPFAKPPLGSLRFAPPQPAEPW	115	EDCLYLNIYTPADLTKSDRLPV
09iEST3_RAT	26	VDTLQGKVLG	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYTPADLTKRDRLPV
09jEST5_RAT	26	VDTMKGKVLG	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYTPADFTKDSRMPV
09kEST4_RAT	26	VDTTKGKVLG	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYTPADFTKNSRLPV
09lSASB_ANAPL	32	VVTNYGSVRG	56	FLGLPFAKPPVGPLRFSEPQPPEPW	120	EDCLYLNVYTPVSTEEQEKLPV
09mESTM_MOUSE	22	VDTPLGRVRG	46	FLGIPFAQAPLGPLRFSAPLPPQPW	108	EDCLTLNIYSPTEITAGDKRPV
11bESTC_DROPS	28	VDLPHGKIRG	46	YESLPYAEPPVGELRFEAPQPYKQQ	101	EDCLTVSIYKPKNTSQSSFPVV
11dESTP_DROME	27	VEITNGKIRG	45	YESIPYAEHPTGALRFEAPQPYSHH	100	EDCLTVSIYKPKKPNRSSFPVV
11eEST6_DROSI	28	VQLPQGKLRG	46	YESIPYAEPPTGDLRFEAPEPYKQK	101	EDCLTVSIYKPKNSKRSTFPVV
11fEST6_DROMA	28	VQLPQGKLRG	46	YESIPYAEPPIGDLRFEAPEPYKQK	101	EDCLTVSIYKPKNSKRSSFPVV
12aACES_TORCA	29	VNTKSGKVMG	51	FLGIPFAEPPVGNMRFRRPEPKKPW	113	EDCLYLNIWVPSPRPKSTTVMV

	Block $1M_4$	
01aPCD_ARTOX	144	PLGFLAGMGDENVWLTDQVEALRWIADNVAAFGGDPNRITLVGQSGGAYSI
02aPNBA_BACSU	145	HLSSFDEAYSDNLGLLDQAAALKWVRENISAFGGDPDNVTVFGESAGGMSI
04aEST1_CAEBR	155	FFSEGTSDAPGNYGLFDQAAALRFVKENIGNFGGDPDDITIWGYSAGAASV
04bEST1_CAEEL	154	FFSEGTSDVQGNWGLFDQAAALEFVKSNIENFGGDPNQITIWGYSAGAASV
06aCRYS_DICDI	171	LGFLCTGLLSGNFGFLDQVMALDWVQENIEVFGGDKNQVTIYGESAGAFSV
06bD2_DICDI	169	LGFMGTDLMHGNYGFLDQIKALEWVYNNIGSFGGNKEMITIWGESAGAFSV
09aEST1_RAT	177	LFSTGDEHSRGNWAHLDQLAALRWVQDNIANFGGNPDSVTIFGESAGGVSV
09bEST1_MESAU	183	FFSTGDEHARGNWGYLDQVAALHWVQQNIASFGGNPGQVTIFGVSAGGTSV
09cEST2_RABIT	157	FFSTGDQHATGNHGYLDQVAALRWVQKNIAHFGGNPGRVTIFGESAGGTSV
09dESTN_MOUSE	177	LFSTGDEHSPGNWAHLDQLAALRWVQDNIANFGGNPDSVTIFGESSGGISV
09eEST1_HUMAN	177	FFSTGDEHSRGNWGHLDQVAALRWVQDNIASFGGNPGSVTIFGESAGGESV
09fEST1_RABIT	151	QYRLGIGGFGFNIDELFLVAVNRWVQDNIANFGGDPGSVTIFGESAGGQSV
09gES10_RAT	177	FFSTGDEHSRGNWGHLDQVAALHWVQDNIANFGGNPGSVTIFGESAGGFSV
09hES22_MOUSE	178	FFSTGDEHSRGNWGHLDQVAALHWVQDNIAKFGGDPGSVTIFGESAGGESV
09iEST3_RAT	177	FFSTGDEHSRGNWGHLDQVAALHWVQDNIDNFGGDPGSVTIFGESAGGESV
09jEST5_RAT	177	FFSTGDEHSRGNWGHLDQVAALHWVQDNIANFGGDPGSVTIFGESAGGFSV
09kEST4_RAT	177	FFSTGDEHSRGNWGHLDQVAALHWVQDNIANFGGDPGSVTIFGESAGGFSV
09lSASB_ANAPL	183	YFSTGDKHARGNWGYLDQVAALQWIQENIIHFRGDPGSVTIFGESAGGVSV
09mESTM_MOUSE	171	FLSTGDKHMPGNRGFLDVVAALRWVQGNIAPFGGDPNCVTIFGNSAGGIIV
11bESTC_DROPS	163	FASTGDADLSGNFGLKDQRLALLWIKQNIASFGGEPENIIVVGHSAGGASV
11dESTP_DROME	162	FASTGDRHLPGNYGLKDQRLALQWIKKNIAHFGGMPDNIVLIGHSAGGASA
11eEST6_DROSI	163	FASTGDRDLPGNYGLKDQRLALKWIKQNIASFGGEPQNVLLIGHSAGGASV
11fEST6_DROMA	163	FASTGDRDLPGNYGLKDQRLALKWIKQNIASFGGEPENILLIGHSAGGASV
12aACES_TORCA	177	LALHGSQEAPGNVGLLDQRMALQWVHDNIQFFGGDPKTVTIFGESAGGASV

Anhang D Konservierte Sequenzblöcke des Sequenzdatensatzes S_1 für das Design der familienspezifischen Primer bestimmt mit Gibbs Sampler.[168]

	Block $1G_1$		Block $1G_2$	
04aEST1_CAEBR	38	FLAIPFAKPPVDNLRFEKPEAPEPW	91	EDCLTLNVIKPK
04bEST1_CAEEL	38	FLAIPFAKPPVDDLRFEKPVAPDPW	91	EDCLTLNIIKPK
06aCRYS_DICDI	54	FYGIPFARPPIDELRYEDPQPPKPW	109	EDCLYLDVFIPR
06bD2_DICDI	52	FYGIPFAQPPVNQLRWENPIDLKPW	107	EDCLYLDVFTPK
08aBAL_RAT	51	FKGIPFATAKTLENPQRHPGWQGTL	98	EDCLYLNIWVPQ
08bBAL_MOUSE	51	FKGIPFATAKTLENPQRHPGWQGTL	98	EDCLYLNIWVPQ
08cBAL_HUMAN	50	FKGIPFAAPTKALENPQPHPGWQGT	98	EDCLYLNIWVPQ
08dBAL_BOVIN	48	FKGIPFAAAPKALEKPERHPGWQGT	96	EDCLYLNIWVPQ
09aEST1_RAT	50	FLGVPFAKPPLGSLRFAPPEPAEPW	114	EDCLYLNIYSPA
09bEST1_MESAU	58	FLGIPFAKPPVGPLRFAPPEPPEPW	120	EDCLYLNIYTPA
09cEST2_RABIT	32	FLGIPFAKPPLGPLRFAPPEPAEAW	94	EDCLYLNIYSPA
09dESTN_MOUSE	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYSPA
09eEST1_HUMAN	50	FLGIPFAKPPLGPLRFTPPQPAEPW	114	EDCLYLNIYTPA
09fEST1_RABIT	32	FLGVPFAKPPLGSLRFAPPQPAESW	96	EDCLYLNIYTPA
09gES10_RAT	50	FLGIPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNVYTPA
09hES22_MOUSE	51	FLGVPFAKPPLGSLRFAPPQPAEPW	115	EDCLYLNIYTPA
09iEST3_RAT	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYTPA
09jEST5_RAT	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYTPA
09kEST4_RAT	50	FLGVPFAKPPLGSLRFAPPQPAEPW	114	EDCLYLNIYTPA
09lSASB_ANAPL	56	FLGLPFAKPPVGPLRFSEPQPPEPW	120	EDCLYLNVYTPV
09mESTM_MOUSE	46	FLGIPFAQAPLGPLRFSAPLPPQPW	108	EDCLTLNIYSPT
10aESTE_MYZPE	52	FLGIPYASPPVQNNRFKEPQPVQPW	104	EDCLFLNVYTPK
11aESTB_DROPS	46	YESLPYAEPPVGDLRFEAPQPYKQQ	101	EDCLTVSVYKPK
11bESTC_DROPS	46	YESLPYAEPPVGELRFEAPQPYKQQ	101	EDCLTVSIYKPK
11cESTA_DROPS	48	YEAIPYAEPPTGELRFEVPKPYKQQ	103	EDCLTVSVYRPK
11dESTP_DROME	45	YESIPYAEHPTGALRFEAPQPYSHH	100	EDCLTVSIYKPK
11eEST6_DROSI	46	YESIPYAEPPTGDLRFEAPEPYKQK	101	EDCLTVSIYKPK
11fEST6_DROMA	46	YESIPYAEPPIGDLRFEAPEPYKQK	101	EDCLTVSIYKPK
11gEST6_DROME	48	YESIPYAEPPTGDLRFEAPEPYKQK	103	EDCLTVSVYKPK
12aACES_TORCA	51	FLGIPFAEPPVGNMRFRRPEPKKPW	113	EDCLYLNIWVPS

	Block $1G_3$		Block $1G_4$	
04aEST1_CAEBR	109	PVLFWVHGGGYEIGS	140	GVIVVTIQYRLGFMGFFSEGTSDAPGNYGLFDQAAALRFVKEN
04bEST1_CAEEL	108	PVLFWIHGGGYEIGS	139	GVIVATVQYRLGFMGFFSEGTSDVQGNWGLFDQAAALEFVKSN
06aCRYS_DICDI	129	PVMVFIPGGAFTQGT	158	SVIVVNVNYRLGVLGFLCTGLLSGNFGFLDQVMALDWVQENIE
06bD2_DICDI	127	PVIVYIPGGAFSVGS	156	SVIVVNINYRLGVLGFMGTDLMHGNYGFLDQIKALEWVYNNIG
08aBAL_RAT	119	PVMVWIYGGAFLMGS	155	NVIVVTFNYRVGPLGFLSTGDANLPGNFGLRDQHMAIAWVKRN
08bBAL_MOUSE	119	PVMVWIYGGAFLMGS	155	NVIVVTFNYRVGPLGFLSTGDANLPGNFGLRDQHMAIAWVKRN
08cBAL_HUMAN	119	PVMIWIYGGAFLMGS	155	NVIVVTFNYRVGPLGFLSTGDANLPGNYGLRDQHMAIAWVKRN
08dBAL_BOVIN	117	PVMIWIYGGAFLMGA	153	NVIVVTFNYRVGPLGFLSTGDSNLPGNYGLWDQHMAIAWVKRN
09aEST1_RAT	134	PVMVWIHGGGLIIGG	162	NVVVVTIQYRLGIWGLFSTGDEHSRGNWAHLDQLAALRWVQDN
09bEST1_MESAU	140	PVMVWIHGGALVMGM	168	DIVIVSIQYRLGILGFFSTGDEHARGNWGYLDQVAALHWVQQN
09cEST2_RABIT	114	PVMVWIHGGGLTMGM	142	DVVVVTIQYRLGVLGFFSTGDQHATGNHGYLDQVAALRWVQKN
09dESTN_MOUSE	134	PVMVWIHGGGLVIGG	162	NVVVVTIQYRLGIWGLFSTGDEHSPGNWAHLDQLAALRWVQDN
09eEST1_HUMAN	134	PVMVWIHGGGLMVGA	162	NVVVVTIQYRLGIWGFFSTGDEHSRGNWGHLDQVAALRWVQDN
09fEST1_RABIT	116	PVMVWIHGGGLMVGG	144	NVVVVTIQYRLGIGGFGFNIDELFLVAVNRWVQDNIANFGGDP
09gES10_RAT	134	PVMVWIHGGGLVVGG	162	NVVVVTIQYRLGIWGFFSTGDEHSRGNWGHLDQVAALHWVQDN
09hES22_MOUSE	135	PVMVWIHGGGLVLGG	163	NVVVVVIQYRLGIWGFFSTGDEHSRGNWGHLDQVAALHWVQDN
09iEST3_RAT	134	PVMVWIHGGGLVLGG	162	NVVVVVIQYRLGIWGFFSTGDEHSRGNWGHLDQVAALHWVQDN
09jEST5_RAT	134	PVMVWIHGGGLTQGG	162	NVVVVAIQYRLGIWGFFSTGDEHSRGNWGHLDQVAALHWVQDN
09kEST4_RAT	134	PVMVWIHGGGMTLGG	162	NVVVVAIQYRLGIWGFFSTGDEHSRGNWGHLDQVAALHWVQDN
09lSASB_ANAPL	140	PVFVWIHGGGLVSGA	168	NVVVVTIQYRLGIAGYFSTGDKHARGNWGYLDQVAALQWIQEN
09mESTM_MOUSE	128	PVMVWIHGGSLRVGS	156	DVVVVTVQYRLGIFGFLSTGDKHMPGNRGFLDVVAALRWVQGN
10aESTE_MYZPE	127	NVIVHIHGGGYYFGE	155	DFVYVSINYRLGVLGFASTGDGVLTGNNGLKDQVAALKWIQQN
11aESTB_DROPS	120	PVVAQIHGGAFMFGG	148	NLILVKISYRLGPLGFVSTGDADLSGNFGLKDQRLALLWIKQN
11bESTC_DROPS	120	PVVAHMHGGAFMFGE	148	KLILVKISYRLGPLGFASTGDADLSGNFGLKDQRLALLWIKQN
11cESTA_DROPS	122	PVVANLHGGAFMFGG	150	SVILVTIGYRLGPLGFVSTGDADLSGNFGLKDQRLALLWIKQN
11dESTP_DROME	119	PVVVLLHGGAFMFGS	147	TLLVVKISYRLGPLGFASTGDRHLPGNYGLKDQRLALQWIKKN
11eEST6_DROSI	120	PVVAHIHGGAFMFGA	148	KFILVKISYRLGPLGFASTGDRDLPGNYGLKDQRLALKWIKQN
11fEST6_DROMA	120	PVVAHIHGGAFMFGA	148	KFILVKISYRLGPLGFASTGDRDLPGNYGLKDQRLALKWIKQN
11gEST6_DROME	122	PVVAHIHGGAFMFGA	150	KFILVKISYRLGPLGFVSTGDRDLPGNYGLKDQRLALKWIKQN
12aACES_TORCA	131	TVMVWIYGGGFYSGS	161	EVVLVSLSYRVGAFGFLALHGSQEAPGNVGLLDQRMALQWVHD

	Block $1G_5$		Block $1G_6$
04aEST1_CAEBR	191	DDITIWGYSAGAASVSQLTMSPYTHDLYSKAIIMSA	258 ECMKKKTLHEIF
04bEST1_CAEEL	190	NQITIWGYSAGAASVSQLTMSPYTRDSYSKAIIMSA	257 ECMKKKSLHEIF
06aCRYS_DICDI	207	NQVTIYGESAGAFSVAAHLSSEKSEGKFHRAILSST	273 DCHRSKSPEEIL
06bD2_DICDI	205	EMITIWGESAGAFSVSAHLTFTYSRQYFNAAISSSS	271 TCLRGKSMDEIL
08aBAL_RAT	206	DNITIFGESAGAASVSLQTLSPYNKGLIRRAISQSG	276 GCLKITDPRALT
08bBAL_MOUSE	206	DNITIFGESAGAASVSLQTLSPYNKGLIRRAISQSG	276 ACLKITDPRALT
08cBAL_HUMAN	206	NNITLFGESAGGASVSLQTLSPYNKGLIRRAISQSG	276 QCLKVTDPRALT
08dBAL_BOVIN	204	DNITLFGESAGGASVSLQTLSPYNKGLIKRAISQSG	274 GCLKITDPRALT
09aEST1_RAT	213	DSVTIFGESAGGVSVSALVLSPLAKNLFHRAISESG	283 QCLRQKTEAELL
09bEST1_MESAU	219	GQVTIFGVSAGGTSVSSLVVSPMSKGLFHGAIMQSG	290 HCLREKTEAEIL
09cEST2_RABIT	193	GRVTIFGESAGGTSVSSHVLSPMSQGLFHGAIMESL	263 RCLRAKSEEEML
09dESTN_MOUSE	213	DSVTIFGESSGGISVSVLVLSPLGKDLFHRAISESG	283 QCLRQKTESELL
09eEST1_HUMAN	213	GSVTIFGESAGGESVSVLVLSPLAKNLFHRAISESG	284 HCLRQKTEEELL
09fEST1_RABIT	187	GSVTIFGESAGGQSVSILLLSPLTKNLFHRAISESG	257 HCLRQKTEEELM
09gES10_RAT	213	GSVTIFGESAGGFSVSALVLSPLAKNLFHRAISESG	283 HCLRQKTEDELL
09hES22_MOUSE	214	GSVTIFGESAGGESVSVLVLSPLAKNLFQRAISESG	284 HCLRQKTEEELL
09iEST3_RAT	213	GSVTIFGESAGGESVSVLVLSPLAKNLFHKAISESG	283 HCLRQKTEEELL
09jEST5_RAT	213	GSVTIFGESAGGFSVSVLVLSPLSKNLYHRAISESG	283 HCLRQKTEEELL
09kEST4_RAT	213	GSVTIFGESAGGFSVSVLVLSPLTKNLFHRAISESG	283 HCLRQKTEEELL
09lSASB_ANAPL	219	GSVTIFGESAGGVSVSALVLSPLAKGLFHKAISESG	288 ECLREKTEAEME
09mESTM_MOUSE	207	NCVTIFGNSAGGIIVSSLLLSPMSAGLFHRAISQSG	277 QCLLQKEGKDLI
10aESTE_MYZPE	206	NSVTITGMSAGASSVHNHLISPMSKGLFNRAIIQSG	276 ECLRSRPAKAIA
11aESTB_DROPS	199	ENILVIGHSAGGGSVHLQVLREDFSKLAKAAISFSG	270 KCLKSKPASEIV
11bESTC_DROPS	199	ENIIVVGHSAGGASVHLQMLREDFAQVAKAGISFGG	270 NCLKSKPAGEIV
11cESTA_DROPS	201	ENILVVGHSAGGASVHLQMLREDFTKVAKAAISFSG	272 KCLKSKPAIEIV
11dESTP_DROME	198	DNIVLIGHSAGGASAHLQLLHEDFKHLAKGAISVSG	269 DCLKSKPASDIV
11eEST6_DROSI	199	QNVLLIGHSAGGASVHLQMLREDFGQLAKAAFSFSG	270 KCLKSKSASELV
11fEST6_DROMA	199	ENILLIGHSAGGASVHLQMLREDFGQLAKAAFSFSG	270 KCLKSKPASELV
11gEST6_DROME	201	QNVLLVGHSAGGASVHLQMLREDFGQLARAAFSFSG	272 KCLKSKPASELV
12aACES_TORCA	213	KTVTIFGESAGGASVGMHILSPGSRDLFRRAILQSG	285 HCLREKKPQELI

Anhang E Konservierte Sequenzblöcke des Sequenzdatensatzes S_2 für das Design der familienspezifischen Primer bestimmt mit MOTIF.[167]

	Block $2M_1$		Block $2M_2$		Block $2M_3$	
01PCD_ARTOX	25	FRGVPYAEPPVGDLRWRAARP	84	EDCLTLNLWTP	102	PVLVWIHGGGLLTG
02PNBA_BACSU	24	WKGIPYAKPPVGQWRFKAPEP	80	EDCLYVNVFAP	98	PVMVWIHGGAFYLG
03EST2_CAEEL	38	YLGIPYAKPPVGELRFKKPVT	95	AGCLTLNVFTP	117	PVMVYIHGGGYELC
04EST1_CAEBR	38	FLAIPFAKPPVDNLRFEKPEA	91	EDCLTLNVIKP	109	PVLFWVHGGGYEIG
06CRYS_DICDI	54	FYGIPFARPPIDELRYEDPQP	109	EDCLYLDVFIP	129	PVMVFIPGGAFTQG
07ESTJ_HELVI	52	FLGVPYAKQPVGELRFKELEP	107	EACIYANIHVP	132	PILVFIHGGGFAFG
09EST1_RAT	50	FLGVPFAKPPLGSLRFAPPEP	114	EDCLYLNIYSP	134	PVMVWIHGGGLIIG
10ESTE_MYZPE	52	FLGIPYASPPVQNNRFKEPQP	104	EDCLFLNVYTP	127	NVIVHIHGGGYYFG
11ESTB_DROPS	46	YESLPYAEPPVGDLRFEAPQP	101	EDCLTVSVYKP	120	PVVAQIHGGAFMFG
12ACES_TORCA	51	FLGIPFAEPPVGNMRFRRPEP	113	EDCLYLNIWVP	131	TVMVWIYGGGFYSG
13LIP2_CANRU	37	FLGIPFAEPPVGTLRFKPPVP	109	EDCLTINVIRP	129	PVMLWIFGGGFELG
14LIP1_GEOCN	43	FKGIPFADPPVGDLRFKHPQP	122	EDCLYLNVFRP	142	PVMVWIYGGAFVFG

	Block $2M_4$		Block $2M_5$	
01PCD_ARTOX	133	LVGISINYRLGPLGFL	154	ENVWLTDQVEALRWIADNVAAFGGDPNRITLVGQSGGAYSI
02PNBA_BACSU	129	VIVVTLNYRLGPFGFL	155	DNLGLLDQAAALKWVRENISAFGGDPDNVTVFGESAGGMSI
03EST2_CAEEL	149	VVVVSINYRLGVFGFL	173	GNFGLWDQTLALKWVQKHISSFGGDPNCVTVFGQSAGGAST
04EST1_CAEBR	141	VIVVTIQYRLGFMGFF	165	GNYGLFDQAAALRFVKENIGNFGGDPDDITIWGYSAGAASV
06CRYS_DICDI	159	VIVVNVNYRLGVLGFL	181	GNFGFLDQVMALDWVQENIEVFGGDKNQVTIYGESAGAFSV
07ESTJ_HELVI	162	VIVITFNYRLNVFGFL	186	GNAGLRDQVTLLRWVQRNAKNFGGDPSDITIAGQSAGASAA
09EST1_RAT	163	VVVVTIQYRLGIWGLF	187	GNWAHLDQLAALRWVQDNIANFGGNPDSVTIFGESAGGVSV
10ESTE_MYZPE	156	FVYVSINYRLGVLGFA	180	GNNGLKDQVAALKWIQQNIVAFGGDPNSVTITGMSAGASSV
11ESTB_DROPS	149	LILVKISYRLGPLGFV	173	GNFGLKDQRLALLWIKQNIASFGGEPENILVIGHSAGGGSV
12ACES_TORCA	162	VVLVSLSYRVGAFGFL	187	GNVGLLDQRMALQWVHDNIQFFGGDPKTVTIFGESAGGASV
13LIP2_CANRU	163	VIHVSMNYRVASWGFL	189	GNAGLHDQRLAMQWVADNIAGFGGDPSKVTIYGESAGSMST
14LIP1_GEOCN	176	VVFVSINYRTGPYGFL	202	TNAGLHDQRKGLEWVSDNIANFGGDPDKVMIFGESAGAMSV

Anhang F Konservierte Sequenzblöcke des Sequenzdatensatzes S_2 für das Design der familienspezifischen Primer bestimmt mit Gibbs Sampler.[168]

	Block $2G_1$		Block $2G_2$	
01PCD_ARTOX	28	VPYAEPPVGDLRWRAARPHAGWTGVRDASAYGPSAPQP	84	EDCLTLNLWTPN
02PNBA_BACSU	27	IPYAKPPVGQWRFKAPEPPEVWEDVLDATAYGPICPQP	80	EDCLYVNVFAPD
03EST2_CAEEL	41	IPYAKPPVGELRFKKPVTVDVWTEIKDCYKYGPACVQT	95	AGCLTLNVFTPR
04EST1_CAEBR	41	IPFAKPPVDNLRFEKPEAPEPWEDVYQATQFRNDCTPH	91	EDCLTLNVIKPK
05EST1_CULPI	34	IPYARAPEGELRFKAPVPPQKWTETLDCTQQCEPCYHF	82	EDSLKINVFAKE
06CRYS_DICDI	57	IPFARPPIDELRYEDPQPPKPWSYVRDGTKQRDQCIQD	109	EDCLYLDVFIPR
07ESTJ_HELVI	55	VPYAKQPVGELRFKELEPLEPWDNILNATNEGPICFQT	107	EACIYANIHVPW
08BAL_RAT	50	IFKGIPFATAKTLENPQRHPGWQGTLKATDFKKRCLQA	98	EDCLYLNIWVPQ
09EST1_RAT	53	VPFAKPPLGSLRFAPPEPAEPWSFVKNTTTYPPMCSQD	114	EDCLYLNIYSPA
10ESTE_MYZPE	55	IPYASPPVQNNRFKEPQPVQPWLGVWNATVPGSACLGI	104	EDCLFLNVYTPK
12ACES_TORCA	54	IPFAEPPVGNMRFRRPEPKKPWSGVWNASTYPNNCQQY	113	EDCLYLNIWVPS
13LIP2_CANRU	40	IPFAEPPVGTLRFKPPVPYSASLNGQQFTSYGPSCMQM	109	EDCLTINVIRPP
14LIP1_GEOCN	46	IPFADPPVGDLRFKHPQPFTGSYQGLKANDFSSACMQL	122	EDCLYLNVFRPA

	Block $2G_3$		Block $2G_4$
01PCD_ARTOX	102	PVLVWIHGGGLLTGSG	133 LVGISINYRLGPLGFL
02PNBA_BACSU	98	PVMVWIHGGAFYLGAG	129 VIVVTLNYRLGPFGFL
03EST2_CAEEL	117	PVMVYIHGGGYELCAS	149 VVVVSINYRLGVFGFL
04EST1_CAEBR	109	PVLFWVHGGGYEIGSG	141 VIVVTIQYRLGFMGFF
05EST1_CULPI	101	PVMLYIYGGGFTEGTS	131 IVLVSFNYRIGALGFL
06CRYS_DICDI	129	PVMVFIPGGAFTQGTG	159 VIVVNVNYRLGVLGFL
07ESTJ_HELVI	132	PILVFIHGGGFAFGSG	162 VIVITFNYRLNVFGFL
08BAL_RAT	119	PVMVWIYGGAFLMGSG	156 VIVVTFNYRVGPLGFL
09EST1_RAT	134	PVMVWIHGGGLIIGGA	163 VVVVTIQYRLGIWGLF
10ESTE_MYZPE	127	NVIVHIHGGGYYFGEG	156 FVYVSINYRLGVLGFA
12ACES_TORCA	131	TVMVWIYGGGFYSGSS	162 VVLVSLSYRVGAFGFL
13LIP2_CANRU	129	PVMLWIFGGGFELGGS	163 VIHVSMNYRVASWGFL
14LIP1_GEOCN	142	PVMVWIYGGAFVFGSS	176 VVFVSINYRTGPYGFL

	Block $2G_5$		Block $2G_6$
01PCD_ARTOX	154	ENVWLTDQVEALRWIADNVAAFGGDPNRITLVGQSGGAYSI	441 WIAFVRTGDPT
02PNBA_BACSU	155	DNLGLLDQAAALKWVRENISAFGGDPDNVTVFGESAGGMSI	436 WITFAKTGNPS
03EST2_CAEEL	173	GNFGLWDQTLALKWVQKHISSFGGDPNCVTVFGQSAGGAST	478 FSNFAKYGNPN
04EST1_CAEBR	165	GNYGLFDQAAALRFVKENIGNFGGDPDDITIWGYSAGAASV	485 VVSFAKTGVPH
05EST1_CULPI	157	GNAGLKDQNLAIRWVLENIAAFGGDPKRVTLAGHSAGAASV	475 FSAFVINGDPN
06CRYS_DICDI	181	GNFGFLDQVMALDWVQENIEVFGGDKNQVTIYGESAGAFSV	477 FVNFIKYSNPS
07ESTJ_HELVI	186	GNAGLRDQVTLLRWVQRNAKNFGGDPSDITIAGQSAGASAA	501 FLNFIKCSQPT
08BAL_RAT	180	GNFGLRDQHMAIAWVKRNIAAFGGDPDNITIFGESAGAASV	488 WTNFAKSGDPN
09EST1_RAT	187	GNWAHLDQLAALRWVQDNIANFGGNPDSVTIFGESAGGVSV	485 WANFARNGNPN
10ESTE_MYZPE	180	GNNGLKDQVAALKWIQQNIVAFGGDPNSVTITGMSAGASSV	495 WATFIKSGVPD
12ACES_TORCA	187	GNVGLLDQRMALQWVHDNIQFFGGDPKTVTIFGESAGGASV	494 WATFAKTGNPN
13LIP2_CANRU	189	GNAGLHDQRLAMQWVADNIAGFGGDPSKVTIYGESAGSMST	485 FIAFANDLDPN
14LIP1_GEOCN	202	TNAGLHDQRKGLEWVSDNIANFGGDPDKVMIFGESAGAMSV	504 FISFANHHDPN

Lebenslauf

Angaben zur Person:

Name:	Markus Fischer
Geburtsdatum:	16. April 1971
Geburtsort:	Ludwigsburg
Staatsangehörigkeit:	Deutsch

Schulbildung:

1977-1981	Grundschule: Pestalozzischule in Ludwigsburg
1981- 1991	Besuch des Friedrich-Schiller-Gymnasium in Ludwigsburg

Akademische Ausbildung:

Oktober 1991 - März 1998	Studium der Chemie an der Universität Stuttgart
März - August 1998	Diplomarbeit am Institut für Technische Biochemie der Universität Stuttgart. Titel der Arbeit: „Protein Engineering der Monooxygenase P450 BM-3“
September 1998	Beginn der Doktorarbeit am Institut für Technische Biochemie.

.