Datenbankgestützte Methoden zur Analyse von Funktions-, Expressions- und Aufreinigungseigenschaften von Proteinen

Von der Fakultät Energie-, Verfahrens- und Biotechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von

Michael Widmann

aus Mutlangen

Hauptberichter:	Prof. Dr. Jürgen Pleiss
Mitberichter:	Prof. Dr. Bernhard Hauer
Tag der mündlichen Prüfung:	22.04.2010

Institut für Technische Biochemie
der Universität Stuttgart

2010

Teile dieser Arbeit wurden bereits veröffentlicht / Parts of this thesis have previously been published:

Widmann, M., Clairo, M., Dippon, J., Pleiss, J., 2008. Analysis of the distribution of functionally relevant rare codons.
BMC Genomics 2008, 9:207

Widmann, M., Trodler P., Pleiss, J., 2010. The isoelectric region of proteins: A systematic analysis.
PLoS ONE 2010, 5 (5)

Widmann, M., Juhl, P., B., Pleiss, J., 2010. Structural classification by the Lipase Engineering Database: a case study of *Candida antarctica* lipase A.
BMC Genomics 2010, 11:123

Widmann, M., Radloff, R., Pleiss, J., 2010. The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations.
BMC Biochemistry 2010, 11:9

The Lipase Engineering Database (LED)
http://www.led.uni-stuttgart.de/
The Thiamine diphosphate dependent Enzyme Engineering Database (TEED)
http://www.teed.uni-stuttgart.de/

Danksagung

Bedanken möchte ich mich bei Herrn Prof. Dr. Jürgen Pleiss für die wissenschaftliche Betreuung, die Überlassung der verschiedenen Themen, sowie die zahlreichen Anregungen und Diskussionen, die entscheidend zum Erfolg dieser Arbeit beigetragen haben.

Bei Herrn Prof. Dr. Schmid und Herrn Prof. Dr. Hauer möchte ich mich für die Möglichkeit bedanken, meine Arbeit am Institut für Technische Biochemie unter ausgezeichneten Arbeitsbedingungen durchführen zu können.

Herrn Prof. Dr. Georg Sprenger danke ich für die Bereitschaft den Prüfungsvorsitz zu übernehmen.

Den Mitarbeitern und Mitarbeiterinnen des Instituts für Technische Biochemie und speziell den Mitgliedern der Bioinformatikgruppe möchte ich für das sehr angenehme Arbeitsklima und die unkomplizierte Zusammenarbeit bei verschiedensten Themen danken.
Zu erwähnen sind hier besonders Dr. Peter Trodler, Dr. Alexander Steudle, Dr. Markus Fischer, Dipl. Biol. (t.o.) Peter B. Juhl, Dr. Demet Sirim, Dipl. Inf. Sascha Rehm, Sven Richter und Florian Wagner.

Ganz besonders möchte ich mich noch bei meiner Mutter, Elisabeth Widmann, für ihre ständige, weit über das Finanzielle hinausgehende Unterstützung während des Studiums und der Promotion bedanken.

Inhaltsverzeichnis

1. Abkürzungsverzeichnis

Aus dem englischen Sprachgebrauch stammende Fachbegriffe, deren direkte Übersetzung keinen Sinn ergeben oder zu einem falschen Verständnis geführt hätte, wurden im Text kursiv dargestellt. Aminosäuren und Nukleinbasen wurden nach den Empfehlungen des NC-IUBMB (*Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*) abgekürzt.

Folgende Abkürzungen wurden in der vorliegenden Arbeit verwendet:

CALA	*Candida antarctica* Lipase A
CALB	*Candida antarctica* Lipase B
CUF	*codon usage frequency*
DBVS	Datenbankverwaltungssystem
DC	Decarboxylase
DWARF	*Data Warehouse for Analysing Protein Families*
E. coli	Escherichia coli
ETL	Extraktion-Transformation-Laden
HMM	*Hidden Markov* Modell
IER	*isoelectric region*
LED	*Lipase Engineering Database*
pI	isoelektrischer Punkt
PP	Pyrophosphat
PYR	Pyrimidin
RCRR	*rare codon rich region*
SNP	s*ingle nucleotide polymorphism*
SQL	*Structured Query Language*
TEED	*Thiamine diphosphate dependent Enzyme Engineering Database*
TH3	Transhydrogenase-dIII-Domäne
ThDP	Thiamindiphosphat
TK	Transketolase
TKC	Transketolase C-terminale-Domäne

2. Zusammenfassung

Die Anzahl an Sequenz- und Strukturinformationen zu Proteinen und Genomen ist gegenwärtig bereits immens umfangreich und wächst stetig an. So umfasste allein die *GenBank* Sequenzdatenbank zum Zeitpunkt dieser Arbeit (Januar 2010) über 100 Millionen Sequenzeinträge und verdoppelt ihren Bestand im Schnitt alle drei Jahre. Die Vielzahl an Sequenz- und Strukturinformationen birgt ein enormes Potenzial, um mittels systematischer Analysen Modelle und Vorhersagen für das Verhalten von Proteinen und gegebenenfalls ganzer Proteinfamilien zu erstellen. Dem gegenüber steht die Art und Weise wie diese Informationen zugänglich sind. Um systematische Analysen ganzer Proteinfamilien durchführen zu können, müssen diese in einem einheitlichen Format vorliegen. Oft stammen die erhältlichen Informationen jedoch aus verschiedenen Quellen und unterscheiden sich in Art und Umfang der verfügbaren Informationen. Die Integration aller für eine Analyse relevanter Daten in einer lokal verfügbaren Datenbank, sowie deren Pflege und Annotierung ist daher entscheidend für die Qualität der Daten und für die Zuverlässigkeit der darauf aufbauenden Analysen.

Das Ziel der in dieser Arbeit durchgeführten systematischen Analysen war die Erstellung von familienspezifischen und allgemeinen Regeln für die Vorhersage von Expressions- und Aufreinigungsparametern von Proteinen. Hierfür wurde zum einen der Einfluss von seltenen Codons auf die Expression von Proteinen untersucht. Zum anderen wurden Modelle erstellt um die Vorhersage des Aufreinigungsverhaltens von Proteinen mittels Ionenaustausch-chromatographie zu beschreiben. Die Proteinfamiliendatenbank der Familie der α/β–Hydrolasen war Teil dieser Analysen und wurde darüber hinaus umfangreich erweitert. Die Klassifizierung und Integration einer neuen Proteinfamilie in diese Datenbank wurde durch systematische Struktur- und Sequenzvergleiche ermöglicht. Zudem wurde für die Familie der Thiamindiphosphat (ThDP)-abhängigen Enzyme eine umfassende Proteinfamiliendatenbank etabliert mit dem Ziel, systematische Analysen dieser diversen Proteinfamilie zu ermöglichen.

In Industrie und Forschung werden oft große Mengen an korrekt gefalteten und aktiven rekombinanten Proteinen benötigt. Der Einfluss der Codon-Nutzung (*codon usage*) auf das Expressionsniveau und den Faltungsweg eines Proteins ist in diesem Zusammenhang seit

langem bekannt. Je nach Organismus werden verschiedene Codons unterschiedlich häufig verwendet. Die Menge an verfügbarer tRNA für ein bestimmtes Codon korreliert dabei mit der Häufigkeit, mit der ein Codon benutzt wird, und bestimmt, wie schnell dieses Codon an den Ribosomen translatiert werden kann. In der Praxis hat sich der Austausch von selten genutzten gegen häufiger genutzte Codons etabliert und geht oft einher mit erhöhten Expressionsmengen. Es gibt jedoch ebenfalls Befunde, die belegen, dass der Austausch seltener Codons negative Folgen, wie die erhöhte Bildung von unlöslichen Proteinen (*inclusion bodies*) und ein Absinken der Enzymaktivität, haben kann. Ein einheitlicher Ansatz zur Bewertung der Relevanz einzelner seltener Codons existiert aufgrund dieser widersprüchlichen Ergebnisse noch nicht. Um ein Modell zur Identifizierung relevanter seltener Codons zu entwickeln, wurden mit computergestützten Methoden systematische Analysen mehrerer Proteinfamilien durchgeführt und mit experimentellen Daten verglichen. Das entwickelte Modell beruht dabei auf dem Ansatz, dass seltene Codons, die relevant für die Geschwindigkeit der Translation beziehungsweise die Ausbildung der Struktur sind, im Laufe der Evolution konserviert wurden. Um Bereiche mit einer erhöhten Anzahl an konservierten seltenen Codons zu detektieren, wurden Multisequenz-Alignments von 18 Proteinfamilien mit insgesamt 157 Proteinen einer Analyse unterzogen. 16 Proteinfamilien wurden dabei der Proteinfamiliendatenbank der α/β–Hydrolasen, der *Lipase Engineering Database* (LED), entnommen. Obwohl die Sequenzähnlichkeit der unterschiedlichen Familien der α/β–Hydrolasen untereinander oft gering ist, gehören alle α/β–Hydrolasen zu derselben *fold* Familie des α/β–Hydrolase-*folds* und sind sich damit strukturell sehr ähnlich. Durch die strukturelle Ähnlichkeit der verschiedenen Proteinfamilien sollte ebenfalls untersucht werden, ob konservierte seltene Codons verschiedener Proteinfamilien gehäuft in bestimmten Strukturmerkmalen des α/β–Hydrolase-*folds* vorkamen. Basierend auf den Sequenzen der Chloramphenicol-Acetyltransferase und des Fettsäure-bindenden Proteins wurden zwei weitere Proteinfamilien manuell erstellt. Diese Proteine waren zuvor das Ziel experimenteller Arbeiten gewesen, die detailliert die Auswirkungen eines Austauschs seltener Codons untersuchten. Bereiche mit einer erhöhten Anzahl seltener Codons wurden dabei als *rare codon rich region* (RCRR) klassifiziert. Es wurden insgesamt 42 RCRR Bereiche bei den untersuchten Proteinfamilien identifiziert. Bei insgesamt 34 RCRR Bereichen war es möglich, diese auf eine 3D Kristallstruktur zu übertragen und damit einem Strukturmerkmal zuzuordnen. Die Vorhersagen wurden anschließend mit experimentellen Befunden verglichen. Dabei zeigte sich, dass sich die Vorhersagen mit experimentell bestätigten Bereichen deckten, in denen seltene Codons gefunden wurden, die relevant für die korrekte

Faltung des jeweiligen Proteins waren. Durch die Einbeziehung der Proteinfamilie der α/β–Hydrolasen war es zusätzlich möglich, die gefundenen Bereiche in Bezug auf ihre Position im gemeinsamen α/β–Hydrolase-*fold* zu untersuchen. Dabei zeigte sich, dass es keine generelle Präferenz von RCRRs für Sekundärstrukturelemente im Vergleich zu *loop* und *linker* Regionen gab. Der allen α/β–Hydrolasen gemeinsame *fold* besitzt zudem keine Region, innerhalb derer RCRRs aller Proteinfamilien gehäuft auftraten.

Das Konzept der Analyse kompletter Proteinfamilien zur Erstellung allgemeingültiger Modelle und Regeln wurde ebenfalls im Rahmen dieser Arbeit auf die elektrostatischen Eigenschaften von Proteinen und deren Einfluss auf die Aufreinigung mittels Ionenaustauschchromatographie angewandt. Die Ionenaustauschchromatographie stellt ein weit verbreitetes und etabliertes Verfahren zur Aufreinigung von Proteinen dar. Sie wird meist als Teil einer mehrstufigen Aufreinigung angewandt und liefert oft bereits als einzelner Aufreinigungsschritt gute Ergebnisse. Da die Bindung der Proteine an die Säule größtenteils durch elektrostatische Wechselwirkungen bedingt ist, hat sich die Vorhersage des isoelektrischen Punktes (pI) für eine erste Einschätzung des benötigten Auftragungs- bzw. Elutions pH-Werts bewährt. Neuere Ergebnisse zeigen jedoch, dass viele Proteine bei pH-Werten binden bzw. eluieren, die mit dem vorhergesagten bzw. gemessenen pI nicht erklärbar sind. Eine mögliche Erklärung für dieses abweichende Verhalten stellt dabei die Verlängerung des nahezu ladungsfreien Bereichs in der Nähe des isoelektrischen Punktes dar. Dieser Bereich wurde in dieser Arbeit aufgrund seiner Ladungsneutralität als *isoelectric region* (IER) bezeichnet. Um die experimentellen Beobachtungen zu erklären und Faktoren für die Vorhersage dieses abweichenden Verhaltens zu identifizieren, wurde eine systematische Analyse der Titrationskurvenprofile von mehreren Proteinfamilien durchgeführt. Ausgewertet wurden dabei über 4600 Proteinen der α/β–Hydrolase Familie sowie über 2600 Sequenzen der Dehydrogenase/Reduktase Familie. Durch die Analyse konnten Faktoren, die mit der Größe der IER korrelieren, bestimmt und ein Modell zur Identifizierung von Proteinen mit großer IER erstellt werden. Zusätzlich wurde das bestehende Datenmodell der *Lipase Engineering Database* (LED) erweitert. Es beinhaltet nun auch Informationen zu den elektrostatischen Eigenschaften der jeweiligen Proteine.

Im Rahmen dieser Arbeit wurde die *Lipase Engineering Database* (LED) umfangreich aktualisiert und die Version 3.0 veröffentlicht. Die Anzahl der enthaltenen Protein- und Sequenzeinträge wurde gesteigert und neue homologe Familien und Superfamilien wurden

integriert. Im Rahmen der neuen Version der LED konnte zudem gezeigt werden, dass die Integration sowohl von Sequenz- als auch von Strukturinformationen oft unerlässlich für die Klassifizierung neuer Proteine sein kann. Die bisher nicht in der LED integrierte Lipase A aus *Candida antarctica* (CALA) besitzt nur eine geringe Sequenzähnlichkeit zu anderen Lipasen. Die Einbeziehung der kürzlich veröffentlichten Proteinstruktur ermöglichte es nun sowohl CALA im Rahmen der LED einzuordnen, als auch verwandte Proteine mit ähnlichen Substratbindungstaschen zu identifizieren. Dies erlaubte Rückschlüsse auf die Funktionsweise von CALA.

Als neue Proteinfamiliendatenbank wurde die *Thiamine diphosphate dependent Enzyme Engineering Database* (TEED) etabliert. Die TEED enthält Sequenzdaten zu 9443 Thiamindiphosphat (ThDP)-abhängigen Enzymen, die in insgesamt 8 Superfamilien und 64 homologe Familien unterteilt wurden. Eine Schwierigkeit bei der Analyse und Klassifizierung der ThDP-abhängigen Enzyme stellt die hohe Sequenzdiversität der verschiedenen Superfamilien dar. Ein gut konserviertes Merkmal aller ThDP-abhängigen Enzyme ist die Bindestelle für ThDP. Diese wird von zwei Proteindomänen gebildet, der Pyrophosphat (PP)- und der Pyrimidin (PYR)-Domäne. Beide Domänen sind strukturell konserviert, besitzen jedoch auf Sequenzebene nur eine sehr geringe Ähnlichkeit zwischen den verschiedenen Familien. Aus diesem Grund müssen beide Domänen für jede Proteinfamilie individuell identifiziert werden. Im Rahmen der Erstellung der TEED wurden daher die PP- und PYR-Domäne für jede Proteinfamilie eindeutig identifiziert und im Datenmodell annotiert. Dies ermöglicht den direkten Zugriff auf die Sequenzen der konservierten Bereiche der diversen Familie der ThDP-abhängigen Enzyme und deren vergleichende Analyse. Die TEED trägt damit zum weiteren Verständnis der familienspezifischen Eigenschaften der ThDP-abhängigen Enzyme bei.

Im Rahmen dieser Arbeit konnte somit ein breites Spektrum an bioinformatischen Methoden erfolgreich auf verschiedene Proteinfamilien angewandt werden. Mittels systematischer Sequenz- und Strukturanalysen konnten Modelle zur Beschreibung experimenteller Beobachtungen bei der Expression und Aufreinigung von Proteinen erstellt werden. Die Aktualisierung der Proteinfamiliendatenbank der α/β–Hydrolasen und die Erstellung der Proteinfamiliendatenbank der ThDP-abhängigen Enzyme erlauben durch die Integration biologisch relevanter Daten weitergehende systematische Analysen dieser Proteinfamilien

.

3. Summary

The amount of information on protein sequences and structures is considerable and is steadily increasing. GenBank contains more than 100 million sequence entries (January 2010) and is doubling its content every three years. The multitude of available protein sequences and structures holds an enormous potential for systematic analyses in order to model or predict the behaviour of single proteins and protein families. In contrast to the huge amount of information however stands its accessibility. A consistent format of the data is necessary for every protein family which is to be used in a systematic analysis. However, the available information is often distributed between different sources and differs in form and complexity. The integration of all necessary information in a local database, as well as its maintenance and annotation is therefore crucial for the reliability of every systematic analysis.

The aim of the systematic analyses performed in this work was the development of family specific and general rules for the prediction of parameters for protein expression and purification. One topic was the influence of synonymous codons on the expression of proteins. Furthermore, a model for the prediction and identification of parameters for protein purification via ion exchange chromatography was devised. The protein family database of α/β hydrolases, which was part of these analyses, was also updated and its content considerably increased. The integration and classification of a new protein family to this database was realized by systematic sequence and structure comparisons. A new protein family database was established for the family of Thiamine diphosphate (ThDP)-dependent enzymes, making a systematic analysis and evaluation of this diverse protein family possible.

Large quantities of functional recombinant proteins are often required for industrial and scientific demands. The influence of the codon usage on the expression level and the folding pathway of proteins is a well known factor in the scientific and industrial community. Depending on the organism, different codons are used with varying usage frequencies. The amount of available tRNAs for a specific codon correlates with the usage frequency of this codon and determines the speed with which it is translated at the ribosome. The exchange of rarely used codons with more frequently used codons often results in increased amounts of expressed protein. There are however reports, showing negative results upon the exchange of rare codons like an increase of inclusion body formation and a loss of enzymatic activity.

There is currently no consistent approach for the assessment of the relevance of a rare codon due to these conflicting observations. In order to develop a model for the identification of relevant rare codons, systematic analyses of protein families were performed and the results compared to experimental observations. The model is based on the influence of rare codons on the translational speed and its influence on protein folding. Rare codons, which contribute to the correct folding pathway of a protein by modulating translational speed are assumed to be conserved in the members of a protein family. In order to identify regions with an increased amount of conserved rare codons, multisequence alignments of 18 protein families with a total of 157 proteins were analysed. 16 protein families were taken from the protein family database of α/β hydrolases, the Lipase Engineering Database (LED). Although the sequence identity between the different families of α/β hydrolases is often low, they share a common fold, the α/β hydrolase fold. This structural similarity of all α/β hydrolase families allowed an analysis of the structural location of conserved rare codons. Two additional protein families were generated manually and contained members of the chloramphenicol-acetyltransferase and the fatty-acid-binding protein family. The proteins had been the focus of previous experimental work, investigating the effects of an exchange of rare codons. Areas with an increased number of rare codons were designated as *rare codon rich regions* (RCRR). A total of 42 RCRRs were identified. 34 RCRRs could be mapped to a protein structure. It could be shown that experimentally determined regions with relevant rare codons were overlapping with RCRRs identified by our model. By investigating the protein family of α/β hydrolases, we analysed the identified RCRRs for a common structural location in the α/β hydrolase fold. We could show that there was no preference of RCRRs for secondary structure elements and that the common α/β hydrolase fold has no region with a high amount of conserved rare codons that is shared by all protein families.

The concept of analysing complete protein families to derive general rules and models was also applied to the electrostatic properties of proteins and their influence on ion exchange chromatography purification methods. Ion exchange chromatography is a well established and widely employed method for protein purification. It is an efficient and scalable technique which often yields good results even as a single purification step and is nearly always included in a multi-step purification protocol. The binding of proteins to ion exchange columns is primarily determined by electrostatic interactions. Therefore, the prediction and determination of a proteins isoelectric point (pI) is a reliable parameter for a first appraisal of the required pH for protein binding or elution. However, recent works have shown that some

proteins have binding or elution pHs that are not in accordance with the predicted or experimentally determined pI values. One possible explanation for this divergent behaviour is the elongation of the area of almost neutral charge in vicinity to the pI. This area has been termed the isoelectric region (IER) in this work. In order to explain experimental observations and identify factors for this behaviour, a systematic analysis of the titration curve profiles of several protein families was performed. Two protein families, the α/β hydrolase family with over 4600 proteins and the dehydrogenase/reductase family with over 2600 proteins were analysed. Factors which correlate with the size of the IER were identified and a model for the identification of proteins with a large IER was developed. Additionally, the current data model of the Lipase Engineering Database (LED) was extended and now includes data on the electrostatic properties of proteins.

In the course of this work, the Lipase Engineering Database was extensively updated and the version 3.0 of the database was released. The number of protein and sequence entries was considerably increased and new homologous families and superfamilies added. The updated release demonstrated the importance of integrating information on sequence and structural data for the classification of novel proteins. The newly introduced superfamily of *Candida antarctica* Lipase A (CALA) has a low sequence identity to other lipases. It was possible to classify CALA in the framework of the LED by comparing the proteins recently published structure to other protein structures of the LED. Structural comparisons also led to the identification of proteins with a similar substrate binding site, leading to insights into the working mechanisms of CALA.

Furthermore, a new database, the Thiamine diphosphate (ThDP)-dependent Enzyme Engineering Database (TEED) was established. The TEED contains sequence entries of 9443 ThDP-dependent enzymes which were assigned to 8 different superfamilies and 63 homologous families. The high sequence diversity between the members of the ThDP-dependent enzyme family makes systematic analyses difficult across family borders. A conserved feature of all ThDP-dependent enzymes is the ThDP binding site. Two common protein domains, the pyrophosphate- (PP) and the pyrimidine- (PYR) domain form the ThDP binding site. Since both domains are only structurally conserved and have low sequence similarity, an individual identification is necessary for each protein family. The PP and PYR domain of each ThDP-dependent protein family was therefore manually identified and

annotated, allowing a direct comparison and systematic analysis of them across family borders.

Within the framework of this thesis, a broad range of bioinformatics methods was successfully applied to different protein families. Using systematic sequence and structure analyses, models for the description of experimental observations in protein expression and purification were developed. By integrating biologically relevant data, the update of the protein family database of α/β hydrolases and the establishment of the protein family database of ThDP-dependent enzymes allow for further systematic analyses of these protein families.

4. Publikationen

Die vorliegende Arbeit umfasst folgende Publikationen:

1. Widmann, M., Clairo, M., Dippon, J., Pleiss, J., 2008. Analysis of the distribution of functionally relevant rare codons. *BMC Genomics* 2008, 9:207
2. Widmann, M., Trodler, P., Pleiss, J. The isoelectric region of proteins: a systematic analysis. *PLoS ONE* 2010, 5 (5)
3. Widmann, M., Juhl, P., B., Pleiss, J. Structural classification by the Lipase Engineering Database: a case study of *Candida antarctica* lipase A *BMC Genomics* 2010, 11:123
4. Widmann, M., Radloff, R., Pleiss, J. The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations. *BMC Biochemistry* 2010, 11:9

The Lipase Engineering Database (LED)
http://www.led.uni-stuttgart.de/

The Thiamine diphosphate dependent Enzyme Engineering Database (TEED)
http://www.teed.uni-stuttgart.de/

5. Einleitung

Mittels systematischer Analysen von Proteinfamilien ist es möglich, Regeln und Modelle für die Vorhersage und Optimierung von Proteineigenschaften aufzustellen. Unabdingbar hierfür sind Proteinfamiliendatenbanken, die alle benötigten Daten in einem einheitlichen und konsistenten Format zur Verfügung stellen. Im Rahmen dieser Arbeit wurden bioinformatische Methoden angewandt, um systematische Analysen von Proteinfamilien durchzuführen und Proteinfamiliendatenbanken zu erstellen. Die vorliegende Einleitung fasst die wesentlichen Grundlagen der systematischen Analysen zusammen. Zusätzlich werden die Proteinfamilien der α/β–Hydrolasen sowie der Thiamindiphosphat (ThDP)-abhängigen Enzyme eingeführt. Diese waren entweder Teil der Analysen oder der Inhalt neuer Proteinfamiliendatenbanken. Abschließend enthält die Einleitung eine kurze Einführung in die Systematik und Technik der benutzten Datenbanksysteme.

5.1. Seltene Codons

5.1.1. Grundlagen der Proteinsynthese

Die Proteine aller Organismen setzen sich aus bis zu 22 proteinogenen Aminosäuren zusammen (Alberts, Johnson et al. 2008). Zahlreiche Aminosäuren werden darüber hinaus nachträglich modifiziert, z. B. durch Phosphorylierung, Methylierung oder Acetylierung (Walsh, Garneau-Tsodikova et al. 2005). Die Codierung der genetischen Information auf DNA-Ebene erfolgt dabei durch Nukleinbasen. Jede Aminosäure wird auf DNA-Ebene durch ein Triplett der Nukleotide Adenin (A), Guanin (G), Thymin (T) und Cytosin (C) codiert (Alberts, Johnson et al. 2008). Ein solches Triplett wird auch als Codon bezeichnet. Die Proteinsynthese (Translation) findet an den Ribosomen statt. Als Vorlage dient dabei ein mRNA-Strang, der zuvor im Laufe der Transkription von der DNA-Vorlage erstellt wurde. Die Nukleinbasen des mRNA-Strangs entsprechen dabei den komplementären Nukleinbasen des DNA-Strangs, mit Ausnahme der Nukleinbase Thymin, die in der mRNA durch die Base Uracil ersetzt ist. Da ein Codon aus drei der vier verschiedenen Basen besteht, gibt es insgesamt $4^3 = 64$ verschiedene Codons. Drei davon codieren nicht für eine Aminosäure, sondern fungieren als sogenannte Stopp-Codons, an denen die Proteinsynthese abgebrochen wird. Dadurch stehen dem Organismus 61 Codons zur Codierung von 20 sogenannten kanonischen Aminosäuren zur Verfügung. Zusätzlich können sie für zwei der nicht-

kanonischen Aminosäuren, Selenocystein (Berry, Tujebajeva et al. 2001) und Pyrrolysin (Fenske, Palm et al. 2003), im Zusammenspiel mit weiteren Faktoren codieren.

5.1.2. Der genetische Code

Da die verbleibenden 61 Codons lediglich 20 verschiedene Aminosäuren codieren und damit mehrere Kombinationen für dieselbe Aminosäure zur Verfügung stehen wird der genetische Code als degeneriert bezeichnet. Für die Aminosäuren Methionin und Tryptophan existiert dabei lediglich ein Codon, für andere (z. B. Arginin) bis zu sechs verschiedene. Alle für die gleiche Aminosäure codierenden Codons werden als synonym bezeichnet. Jedes Codon wird an den Ribosomen durch eine passende tRNA erkannt. Diese besitzt ein zu dem jeweiligen Codon komplementäres Triplett an Nukleotiden und ist mit der dadurch codierten Aminosäure beladen. Manche tRNAs können dabei verschiedene Codons erkennen und die entsprechende Aminosäure einbauen. Dafür ist meist die 3. Base des Tripletts auf der mRNA verantwortlich. Diese wird als wackelnde (*wobble*) Base bezeichnet, da die Base an dieser Stelle in der Lage ist, mit verschiedenen Basen zu interagieren (Crick 1966) (Abbildung 5-1). Als Besonderheit kann das komplementäre Triplett der tRNA ein fünftes Nukleotid, Inosin (I) enthalten welches in der Lage ist mit drei verschiedenen Nukleotiden der mRNA zu interagieren.

***wobble* - Abweichung**	
Base im Anticodon der tRNA	**Base an der 3. Position des Codons**
C	G
A	U
U	A oder G
G	U oder C
I	U, C oder A

Abbildung 5-1: Übersicht der möglichen Paarungen an der 3. Position des mRNA-Codons.

Zur Kodierung einer bestimmten Aminosäure werden die verschiedenen, synonymen Codons jedoch nicht gleichmäßig genutzt. Je nach Spezies werden unterschiedliche Codons bevorzugt für die Codierung der jeweiligen Aminosäure eingesetzt. (Grantham, Gautier et al. 1980). Dadurch gibt es für eine Aminosäure Codons, die im Vergleich zu anderen, synonymen Codons häufiger oder weniger häufig benutzt werden. Es konnte gezeigt werden, dass die Menge an verfügbarer tRNA mit der Frequenz eines Codons korreliert (Ikemura 1985). Für viele Organismen existieren Tabellen mit den Angaben, wie hoch die Frequenz eines jeweiligen Codons ist (Nakamura, Gojobori et al. 2000). Um die Werte für eine solche Tabelle zu bestimmen, wird die Anzahl aller Codons in den für Proteine codierenden DNA-

Bereichen ermittelt und der Quotient aus der Anzahl des jeweiligen Codons und der Gesamtzahl aller Codons gebildet. Das Ergebnis ist die c*odon usage frequency* (CUF), die angibt, welchen Anteil (in Promille) ein Codon an der Gesamtheit der Codons hat (Abbildung 5-2).

```
Escherichia coli CFT073 [gbbct]: 5379 CDS's (1581056 codons)
fields: [triplet] [frequency: per thousand] ([number])

UUU 23.2( 36738)  UCU  8.7( 13723)  UAU 16.5( 26077)  UGU  5.5(  8732)
UUC 16.9( 26655)  UCC  8.9( 14004)  UAC 12.1( 19204)  UGC  6.9( 10920)
UUA 13.9( 22000)  UCA  7.8( 12367)  UAA  2.0(  3217)  UGA  1.1(  1739)
UUG 14.0( 22187)  UCG  8.7( 13729)  UAG  0.3(   426)  UGG 15.2( 23955)

CUU 11.7( 18499)  CCU  7.3( 11563)  CAU 13.6( 21480)  CGU 20.3( 32051)
CUC 11.0( 17321)  CCC  5.8(  9154)  CAC  9.8( 15490)  CGC 21.0( 33132)
CUA  4.0(  6256)  CCA  8.5( 13508)  CAA 15.0( 23778)  CGA  3.9(  6132)
CUG 50.9( 80431)  CCG 21.8( 34497)  CAG 29.5( 46720)  CGG  6.3(  9988)

AUU 29.8( 47046)  ACU  9.1( 14447)  AAU 18.6( 29426)  AGU  9.5( 14953)
AUC 24.2( 38317)  ACC 22.8( 35978)  AAC 21.4( 33762)  AGC 16.0( 25265)
AUA  5.4(  8607)  ACA  8.2( 12981)  AAA 33.2( 52512)  AGA  2.9(  4660)
AUG 27.0( 42762)  ACG 14.8( 23419)  AAG 10.7( 16855)  AGG  1.9(  3041)

GUU 18.5( 29188)  GCU 15.6( 24683)  GAU 32.1( 50786)  GGU 24.4( 38610)
GUC 15.1( 23896)  GCC 25.1( 39754)  GAC 18.6( 29331)  GGC 27.9( 44073)
GUA 11.1( 17622)  GCA 20.6( 32557)  GAA 38.2( 60419)  GGA  9.0( 14209)
GUG 25.5( 40384)  GCG 31.7( 50057)  GAG 17.7( 27971)  GGG 11.3( 17812)
```

Abbildung 5-2: Darstellung der c*odon usage* von *E. coli CFT073*. Nutzungsfrequenzen sind in Promille angegeben. Die Anzahl des jeweiligen Codons ist in Klammern angegeben. Quelle (Nakamura, Gojobori et al. 2000).

Anzumerken ist in diesem Zusammenhang, dass von vielen Organismen die Sequenzierung des kompletten Genoms noch nicht abgeschlossen ist bzw. nicht sämtliche codierenden Bereiche als solche identifiziert wurden. Daher können sich die angegebenen Werte durch neue Erkenntnisse über den jeweiligen Organismus verändern.

5.1.3. Vorkommen und Funktion in biologischen Systemen

Eine allgemeingültige Definition, ab welcher c*odon usage frequency* ein Codon als selten zu bezeichnen ist, existiert nicht. Die Bezeichnung wird daher auf unterschiedliche Codons angewandt und kann, je nach Quelle, abweichen. Es ist jedoch zu beobachten, dass die Anzahl der Genkopien der tRNA, die das jeweilige Codon am Ribosom erkennt, mit abnehmender Frequenz des Codons ebenfalls abnimmt (Hannig and Makrides 1998). Es gibt sogar Codons,

die kein entsprechendes tRNA-Gegenstück besitzen. Diese werden allein durch *wobble* tRNAs bedient, wobei es sich meist um sehr selten benutzte Codons handelt (Mattes 1993).
Gene, die einen hohen Anteil seltener Codons aufweisen, werden durch den Mangel an passenden tRNAs langsamer translatiert als Gene mit weniger seltenen Codons (Pedersen 1984). Die Rolle und die Bedeutung von seltenen Codons sind dabei bis heute nicht eindeutig geklärt. Es wurden etliche Experimente und Untersuchungen zu diesem Thema durchgeführt, meist indem Codons einer DNA-Sequenz gegen synonyme Codons ausgetauscht wurden. Die Mutanten wurden daraufhin auf Veränderungen im Expressionsspiegel und dem Faltungsverhalten des Proteins untersucht. Die Ergebnisse einiger Gruppen deuten darauf hin, dass die höhere oder niedrigere Geschwindigkeit, mit der verschieden häufige Codons translatiert werden, einen Einfluss auf die Faltung des jeweiligen Proteins (Purvis, Bettany et al. 1987) oder die Halbwertszeit der jeweiligen mRNA (Hoekema, Kastelein et al. 1987) hat. So konnte gezeigt werden, dass α-Helices oft von häufiger vorkommenden Codons codiert werden, β-Faltblätter hingegen eher von selteneren Codons (Thanaraj and Argos 1996). In einer anderen Arbeit wurde gezielt ein in *E .coli* vorhandenes Protein untersucht und drei seltene Codons durch synonyme Codons substituiert. Das Ergebnis war eine Mutante, die sich durch eine erheblich schlechtere Expressionsleistung und eine gesteigerte Menge an falsch gefalteten Proteinen auszeichnete (Cortazzo, Cervenansky et al. 2002). Andere Ergebnisse belegen hingegen, dass der Austausch von seltenen Codons zu einer deutlichen Steigerung der Expressionsleistung führen kann. So konnte z. B. die Expression eines Gens aus *Plasmodium falciparum* in *E. coli* durch die Optimierung der c*odon usage* erheblich gesteigert werden. Ein großer Anteil des erhaltenen Proteins war jedoch falsch gefaltet oder lag unlöslich in Form von *inclusion bodies* vor (Yadava and Ockenhouse 2003). Eine andere Gruppe untersuchte ebenfalls die Expression eines Gens aus *Plasmodium falciparum* in *E. coli*. Bei dieser Gruppe konnte die Expressionsleistung durch Codon-Optimierung um den Faktor 3 gesteigert werden. Dabei zeigte sich, dass die Originalsequenz aus *Plasmodium falciparum* nicht nur schlechter exprimiert wurde, sondern diese auch zusätzlich das Wachstum der Bakterien merklich beeinträchtigte. Eine Supplementation durch zusätzliche tRNA-Gene zeigte hingegen keinerlei Wirkung. Diese Gruppe konnte keine nachteilige Wirkung der optimierten Sequenz auf die Faltung des Proteins feststellen (Zhou, Schnake et al. 2004). Ein anderer Ansatz besteht darin, nicht nur einzelne Codons zu ersetzen, sondern ein Gen komplett neu zu synthetisieren. Dabei wurde die c*odon usage* derjenigen eines stark exprimierten Gens in *E. coli* nachempfunden und nur häufig vorkommende Codons benutzt. Das Ergebnis war eine deutlich gesteigerte Expressionsleistung, wobei die Löslichkeit des Proteins nicht

beeinträchtigt wurde (Hale and Thompson 1998). Auch größer angelegte systematische Analysen wurden durchgeführt. So konnte bei der Untersuchung von über 700 Strukturen mit der zugehörigen mRNA-Sequenz kein Zusammenhang zwischen Sekundärstrukturelementen und dem Vorkommen von seltenen Codons festgestellt werden (Brunak and Engelbrecht 1996).
Trotz dieser unterschiedlichen Ergebnisse herrscht weitgehend Einigkeit über folgende Fakten:

- Seltene Codons führen durch den Mangel an entsprechender tRNA zu einer Verlangsamung der Translation an den Ribosomen. Dies kann zu einer Verringerung der Halbwertszeit der mRNA führen oder die Translation an den Ribosomen frühzeitig terminieren (Garel, Chavancy et al. 1981; Curran and Yarus 1989; Mattes 1993; Thanaraj and Argos 1996; Musto, Romero et al. 2003).
- Die Faltung eines Proteins wird durch die von seltenen Codons erzeugten Pausen an den Ribosomen beeinflusst. Die Eliminierung von Pausen durch den Austausch gegen häufiger genutzte Codons kann zu Fehlfaltungen führen (Thanaraj and Argos 1996; Li, Luo et al. 2003).
- Der Austausch von seltenen Codons kann zu einer deutlichen Steigerung der Expressionsleistung führen. In manchen Fällen geht diese gesteigerte Expression jedoch mit einem Anstieg an falsch gefaltetem Protein einher (Cortazzo, Cervenansky et al. 2002; Yadava and Ockenhouse 2003).
- Die Supplementation von tRNA-Kopien selten genutzter Codons kann den gleichen Effekt erzielen wie der Austausch des seltenen Codons gegen ein häufiger genutztes, in manchen Fällen scheint dieser Mechanismus jedoch nicht zu funktionieren (Hua, Wang et al. 1994; Mattes 2001; Zhou, Schnake et al. 2004).

Es gibt also offensichtlich Regulationsmechanismen, die sich sowohl bei der Translation als auch bei der Faltung der entstehenden Aminosäurekette auf seltene Codons stützen. Den vorliegenden Ergebnissen nach ist dieser Mechanismus jedoch nicht generell auf jede Sequenz und jeden Organismus übertragbar. So scheint es Sequenzregionen zu geben, bei denen sich ein Austausch von synonymen Codons stark bemerkbar macht, während ein Austausch an anderer Stelle keinen offensichtlichen Effekt nach sich zieht. Es ist anzunehmen, das seltene Codons, die für die korrekte Faltung und Expression eines Proteins

relevant sind, aufgrund ihrer Bedeutung ein konserviertes Merkmal in einer Proteinfamilie darstellen.

5.2. Elektrostatisches Potenzial

Das elektrostatische Potenzial eines Proteins entsteht durch polarisierte oder geladene Aminosäuren und gebundene Ionen. Die Elektrostatik spielt dabei sowohl bei der Proteinfaltung, der Wechselwirkung von Proteinen miteinander als auch der Funktion des jeweiligen Proteins eine entscheidende Rolle (Warshel, Naray-Szabo et al. 1989). Zur Berechnung der Gesamtladung sowie der Ladungsverteilung müssen dabei die Protonierungszustände der einzelnen Seitenketten bestimmt werden. Ausgehend von einer Proteinsequenz werden dazu die bekannten pKa-Standardwerte der einzelnen Aminosäuren genutzt. Diese Standardwerte weichen teilweise von den realen Werten ab, dies gilt vor allem für katalytisch aktive Aminosäuren (Warshel 1981). Die Abweichung der Protonierungszustände ergibt sich dabei aus der Sensitivität der einzelnen pKa-Werte zur jeweiligen elektrostatischen Umgebung. Liegen Proteinkristallstrukturen vor, kann mittels der Poisson-Boltzmann Gleichung der pKa-Wert jeder Aminosäure berechnet werden. Da die in der vorliegenden Arbeit systematischen Analysen kompletter Proteinfamilien vorrangig auf Sequenzdaten gestützt waren, musste die Berechnung der Gesamtladung mittels sequenzbasierter Methoden ohne Berücksichtigung von pKa-Abweichungen erfolgen.

5.3. Ionenaustauschchromatographie

Ionenaustauschchromatographie ist eine der meist benutzten Chromatographietechniken zur Aufreinigung von Proteinen und Biomolekülen. Dies liegt sowohl an dem hohen Auflösungs- und Aufnahmevermögen als auch an der breiten Anwendbarkeit der Methode (Sheehan and FitzGerald 1996; Palekar, Vasudevan et al. 2000; Healthcare 2004). Das Wirkungsprinzip der Ionenaustauschchromatographie beruht auf der reversiblen Bindung gelöster, geladener Moleküle an immobilisierte Ladungsträger mit entgegengesetzter Ladung (Hallgren, Kalman et al. 2000). Dabei verdrängen Moleküle, die eine stärkere elektrostatische Wechselwirkung mit den immobilisierten Ladungsträgern haben, Moleküle mit schwächeren Wechselwirkungen. Als Ladungsträger werden verschiedene chemische Gruppen verwendet, die sich sowohl in ihrer Ladungsstärke, dem pH-Bereich, an dem sie optimale

Aufreinigungseigenschaften zeigen, als auch der Art der getragenen Ladung unterscheiden. Bestehen die immobilisierten Gruppen aus positiv geladenen Ionen und binden diese dadurch negativ geladene Gegenionen, spricht man von einem Anionenaustauscher. Im Fall von immobilisierten Gruppen mit negativer Ladung, die positiv geladene Gegenionen binden, von einem Kationenaustauscher. Der prinzipielle Ablauf einer Ionenaustauschchromatographie kann dabei in vier Schritte unterteilt werden. Zuerst wird die Ionenaustauschsäule äquilibriert, d. h. die Säule wird durch einen Puffer auf den gewünschten pH-Wert und die Ionenstärke eingestellt. Die immobilisierten Ladungsgruppen werden mit leicht austauschbaren Gegenionen (z. B. Natrium oder Chlorid) beladen. Im zweiten Schritt wird das aufzureinigende Protein zusammen mit eventuell vorhandenen Verunreinigungen aufgetragen. Dabei verdrängen die geladenen Proteine die gebundenen Gegenionen und binden je nach Ladung reversibel und unterschiedlich stark an die immobilisierten Ladungsträger. Der dritte Schritt besteht in der stufenweisen Ablösung der gebundenen Proteine. Dies kann durch eine Erhöhung der Salzkonzentration erfolgen, wodurch nach und nach die gebundenen Proteine gelöst werden, beginnend mit den am schwächsten bindenden Proteinen (Healthcare 2004). Ein anderer Ansatz besteht in der Erhöhung bzw. Erniedrigung des pH-Werts (Pabst, Carta et al. 2008). Je weiter die Proteine sich ihrem isoelektrischen Punkt annähern, desto schwächer wird ihre Bindung. Der vierte Schritt besteht aus der Entfernung eventuell vorhandener Verunreinigungen und der erneuten Äquilibrierung der Säule (Sheehan and O'Sullivan 2001; Healthcare 2004).

5.4. Systematische Analyse von Proteinfamilien

5.4.1. Aufbau von Datenbanken

Die ersten computergestützten Datenbanksysteme entstanden Ende der 1960er Jahre (Codd 1970). Unabhängig vom benutzten Datenmodell (hierarchisch, relational oder objektorientiert) besteht ein Datenbanksystem immer aus zwei Komponenten, den eigentlichen Daten und dem Datenbankverwaltungssystem (DBVS). Das DBVS regelt und verwaltet alle Zugriffe auf die Daten. Ein Zugriff auf die gespeicherten Daten findet also nicht direkt durch das jeweils genutzte Programm statt, sondern immer über das zwischengeschaltete DBVS. Die zur Speicherung der eigentlichen Daten verwendete Datenstruktur wird vom DBVS interpretiert und ist dadurch unabhängig von den verwendeten Anwendungsprogrammen. Zusätzlich sorgt das DBVS für die Einhaltung der referenziellen

Integrität der Daten im laufenden Betrieb. Das Auslesen und Manipulieren der Daten erfolgte in der vorliegenden Arbeit über die *Structured Query Language* (SQL). Diese Abfragesprache stellt die Schnittstelle zwischen dem Datenbanksystem und den Anwendungsprogrammen dar (Jamison 2003) (Abbildung 5-3). SQL wurde in den 1970er Jahren bei IBM in San Jose, Kalifornien, entwickelt. Da immer mehr Firmen SQL-basierte DBVS entwickelten, hat sich die Sprache mittlerweile als Standard für relationale DBVS etabliert. Große bekannte DBVS, die SQL nutzen, sind z. B. Firebird, MySQL und Oracle, aber auch Heimanwendungen, wie z. B. Microsoft Access.

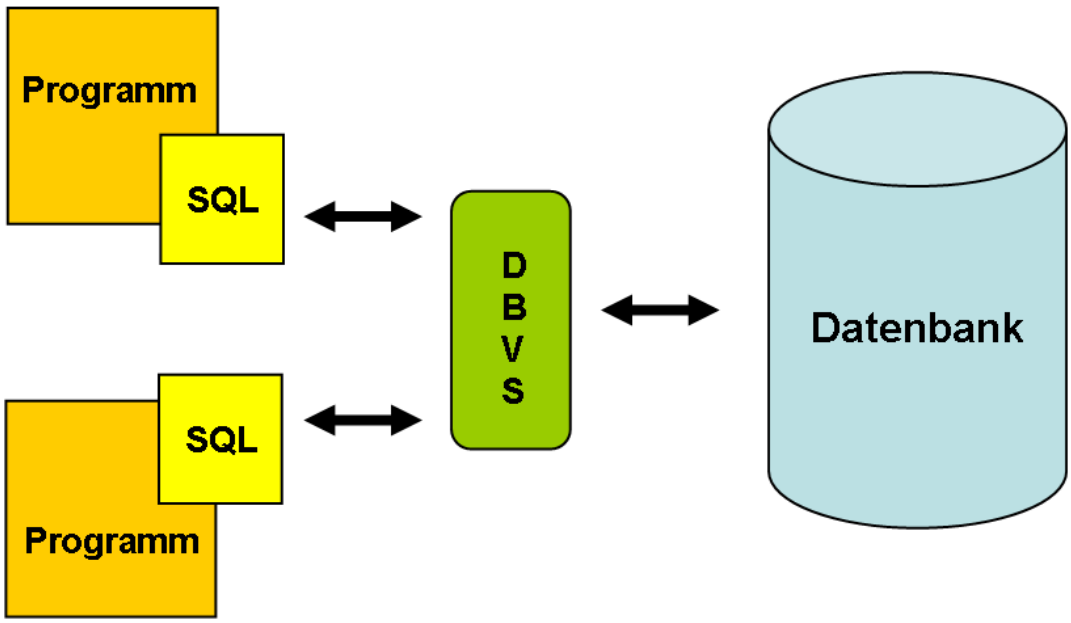


Abbildung 5-3: Schematischer Aufbau eines Datenbanksystems und der darauf zugreifenden Programme. Externe Programme greifen nicht direkt auf die Datenbank zu, sondern immer über das zwischengeschaltete Datenbankverwaltungssystem (DBVS).

5.4.2. Relationale Datenbankmodelle

Alle in der vorliegenden Arbeit benutzten Datenbanken benutzen ein relationales Datenmodell. Bei diesem Modell werden alle Daten in Form von Tabellen gespeichert (Codd 1970). Jede dieser Tabellen ist zweidimensional und besitzt einen Primärschlüssel *(primary key*), dessen Werte in dieser Tabelle für jeden Eintrag eindeutig sind. Zusätzlich können pro Tabelle mehrere Fremdschlüssel (*foreign keys*) definiert werden, die auf die Werte der Primärschlüssel anderer Tabellen verweisen. Dadurch kann z. B. gewährleistet werden, dass jedem Protein eindeutig eine DNA-Sequenz zugeordnet ist. Auf diese Art und Weise können

die verschiedenen Tabellen und deren Inhalte eindeutig miteinander verknüpft und abgefragt werden. So kann z. B. eine Tabelle mit Proteinsequenzen angelegt werden, die mit der Tabelle, in der die DNA-Sequenzen abgelegt sind, verknüpft ist. Die Reihenfolge in der die Daten in den Spalten und Zeilen der Tabelle abgelegt sind, ist dabei irrelevant. Durch die Verknüpfung der verschiedenen Tabellen untereinander kann gewährleistet werden, dass sich Änderungen auf den gesamten Datenbestand auswirken. So würde z. B. das Löschen eines Proteineintrags aus dem obigen Beispiel zu einem DNA-Eintrag führen, der keiner Proteinsequenz mehr zugeordnet ist. In einer relationalen Datenbank ist dies nicht möglich. Es muss also ebenfalls die DNA-Sequenz gelöscht oder einem anderen Eintrag zugeordnet werden, bevor der Proteineintrag gelöscht werden kann.

5.4.3. Data Warehouse

Ein *Data Warehouse* ist eine Datenbank, deren Inhalt größtenteils durch das Zusammenführen und Kopieren von Daten aus anderen Quellen entsteht (Inmon 2002). Mittlerweile sind unzählige biologische Datenbanken im Internet zugänglich. Die gesuchten Daten sind jedoch oft auf verschiedene Datenbanken verteilt die sich in ihren Datenstrukturen unterscheiden. Hinzu kommt, dass die Datenbanken meist individuelle Zugriffsoberflächen benutzen, die nicht ohne Weiteres aufeinander übertragbar sind. Zusätzlich sinkt die Leistung des Systems und steigt der Zeitaufwand für einen Zugriff, je mehr Daten aus verschiedenen Datenbanken benötigt werden. Die Lösung dieses Problems liegt im Zusammenführen aller für die jeweilige Aufgabenstellung benötigten Daten in einer neuen Datenbank, dem sogenannten *Data Warehouse*. Das Extrahieren der Daten und die Umwandlung in ein einheitliches Format erfolgen dabei in einem sogenannten Extraktion-Transformation-Laden (ETL)–Prozess (Inmon 2002). Die auf diese Art gewonnenen Daten bilden nun die neue Datenbank und benötigen zusätzlich ein Datenbankverwaltungssystem, um ein vollständiges Datenbank-system zu bilden (Abbildung 5-4). Die Vorteile sind erheblich beschleunigte Zugriffszeiten, ein einheitliches Datenformat und eine lokale Verfügbarkeit, die bei Datenbanken im Internet nicht immer gegeben ist. Durch die Automatisierung des ETL-Prozesses kann die neue Datenbank auf dem neuesten Stand gehalten werden.

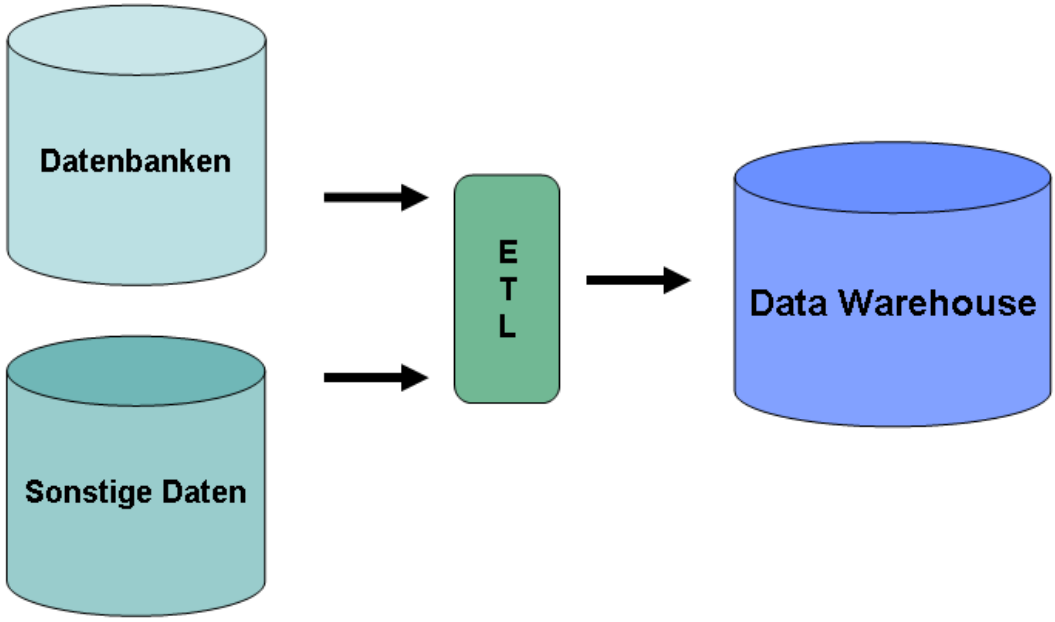


Abbildung 5-4: Schematische Darstellung der Erstellung eines *Data Warehouse*. Daten aus verschiedenen Quellen (hier aus verschiedenen Datenbanken und sonstigen Quellen) werden durch den ETL–Prozess extrahiert, aufbereitet und in die neue Datenbank integriert.

5.4.4. Integration biologischer Daten

Die Anzahl an sequenzierten Proteinen und Genomen, bedingt durch zahlreiche Genomsequenzierungsprojekte, wächst exponentiell an (Gaasterland 1998; Kersey, Bower et al. 2005; Benson, Karsch-Mizrachi et al. 2009). Die aktuelle Version der *GenBank* umfasst über 100 Millionen Sequenzeinträge (Stand 2010) und verdoppelt die Zahl ihrer Einträge alle 3 Jahre (Benson, Karsch-Mizrachi et al. 2009). Viele der verfügbaren Sequenzen stammen aus automatisierten Prozessen. Annotationen zu Eigenschaften und Funktionen werden meist ebenfalls automatisiert über Sequenzvergleiche abgeleitet und zugewiesen. Um weitergehende Aussagen über die Funktion und Eigenschaften von Proteinen treffen zu können, ist es hilfreich diese in biologisch relevante Familien zu klassifizieren. Der Großteil aller verfügbaren Sequenz- und Strukturinformationen befindet sich in öffentlich zugänglichen Sequenz- und Strukturdatenbanken wie der bereits erwähnten *GenBank* (Benson, Karsch-Mizrachi et al. 2009) oder der *Protein Data Bank* (Berman, Westbrook et al. 2000). Neben der Gesamtstruktur sind oft vor allem einzelne Domänen eines Proteins für dessen Aktivität und Zugehörigkeit zu einer Proteinfamilie von Bedeutung (Duggleby 2006). Eine Datenbank die umfassende Informationen über Proteinfamilien und Proteindomänen zur Verfügung stellt ist

InterPro (Hunter, Apweiler et al. 2009). Proteine in der *InterPro* Datenbank werden anhand konservierter Bereiche und funktioneller Abschnitte eingeteilt. *InterProScan* nutzt bereits klassifizierte Abschnitte als Grundlage zur Erstellung von Suchmustern zur Erkennung und Identifizierung von Proteindomänen in unbekannten Proteinen (Quevillon, Silventoinen et al. 2005). Die große Anzahl an Daten und Datenbanken und nicht zuletzt die zunehmende Diversität der Daten stellen jedoch ein nicht unerhebliches Problem dar. Das primäre Ziel der großen Datenbanken ist es, einen möglichst breiten Sequenzraum abzudecken und so viele Informationen wie möglich zu integrieren. Dazu kommt, dass in vielen Datenbanken (z. B. *GenBank*) der Umfang und die Art der zu einem Proteineintrag verfügbaren Informationen sehr heterogen sind, da diese vom Autor des jeweiligen Sequenzeintrags abhängen. Dies erschwert den systematischen Vergleich von Proteineinträgen aus verschiedenen Datenbanken und oft auch aus derselben Datenbank. Der oben vorgestellte Ansatz des *Data Warehouse* Systems ermöglicht die Speicherung der gewünschten Daten in einem einheitlichen und damit vergleichbaren Datenmodell. Das am Institut für Technische Biochemie entwickelte Datenbanksystem DWARF (*Data Warehouse for Analysing Protein Families*) (Fischer, Thai et al. 2006) ermöglicht es, Sequenz- und Strukturdaten in ein lokales System zu integrieren und in einer einheitlichen Datenstruktur zu speichern. Dadurch wird die systematische Analyse von definierten Proteinfamilien ermöglicht, was bereits anhand zahlreicher Datenbanken gezeigt werden konnte. Datenbanksysteme, die erfolgreich mittels des DWARF-Ansatzes erstellt wurden, sind die *Lipase Engineering Database* (Fischer and Pleiss 2003), die *Cytochrome P450 Engineering Database* (Fischer, Knoll et al. 2007), die *Medium-Chain Dehydrogenase/Reductase Engineering Database* (Knoll and Pleiss 2008) und die *PHA Depolymerase Engineering Database* (Knoll, Hamm et al. 2009).

5.5. α/β-Hydrolasen

Die α/β-Hydrolasen sind eine große Familie strukturell verwandter Proteine. Die darin enthaltenen Proteine katalysieren eine Vielzahl von unterschiedlichen chemischen Reaktionen. Mitglieder dieser Familie sind unter anderem Lipasen, Esterasen, Acetylcholinesterasen, Phospholipasen, Cutinasen, Dehalogenasen und Epoxidhydrolasen (Holmquist 2000; Barth, Fischer et al. 2004). Das verbindende Element der verschiedenen Vertreter der α/β-Hydrolasen ist neben der hydrolytischen Spaltung verschiedenster Substrate die gemeinsame Struktur. Obwohl die Sequenzidentität verschiedener α/β-Hydrolasen

teilweise sehr gering ist, nehmen diese dennoch ein gemeinsames Faltungsmuster ein (Ollis, Cheah et al. 1992). Ein weiteres strukturell konserviertes Merkmal ist das Vorhandensein von Aminosäuren die eine katalytische Triade bilden. Diese setzt sich aus einem Nukleophil, einer Säure sowie Histidin zusammen. Als nukleophiles Element dient zumeist Serin, es wurden jedoch auch Aspartat und Cystein beobachtet. Als katalytische Säure fungiert Aspartat oder Glutamat (Schrag, Li et al. 1991).

5.5.1. Struktur der α/β-Hydrolasen

Der α/β-Hydrolase-*fold* ist ein Faltungsmuster das allen α/β-Hydrolasen gemein ist. Es besitzt einen Kern, der aus einem zentralen, überwiegend parallelen β-Faltblatt besteht. Das β-Faltblatt besitzt dabei eine linkshändige Verdrillung, sodass der erste und letzte β-Strang sich in einem Winkel von etwa 90° kreuzen (Ollis, Cheah et al. 1992). Er gehört zum *doubly wound* α/β-*superfold* (Richardson 1981; Orengo, Jones et al. 1994). α-Helices sind auf beiden Seiten des β-Faltblattes angelagert und verbinden die β–Stränge 3-8 miteinander (Abbildung 5-5). Obwohl diese Topologie stets bei allen α/β-Hydrolasen erhalten bleibt, gibt es verschieden starke Unterschiede zwischen den Enzymen, vor allem in der Position und der Länge der α-Helices. Dabei stellt die zwischen den β–Strängen 6 und 7 lokalisierte Helix D einen der am meisten abgewandelten Teile des α/β-Hydrolase-*folds* dar. In dieser Region befindet sich bei manchen α/β-Hydrolasen das sogenannte *cap*, das einen Aufsatz auf den gemeinsamen α/β-Hydrolase-*fold* darstellt. Die Aufgaben des *caps* können vielfältig sein (Wei, Contreras et al. 1999). Sie reichen von der Abschirmung der *active site* vor dem Lösungsmittel (Vincent, Charnock et al. 2003) bis hin zur Beteiligung an der Bildung der Substratbindungstasche (Ericsson, Kasrayan et al. 2008).

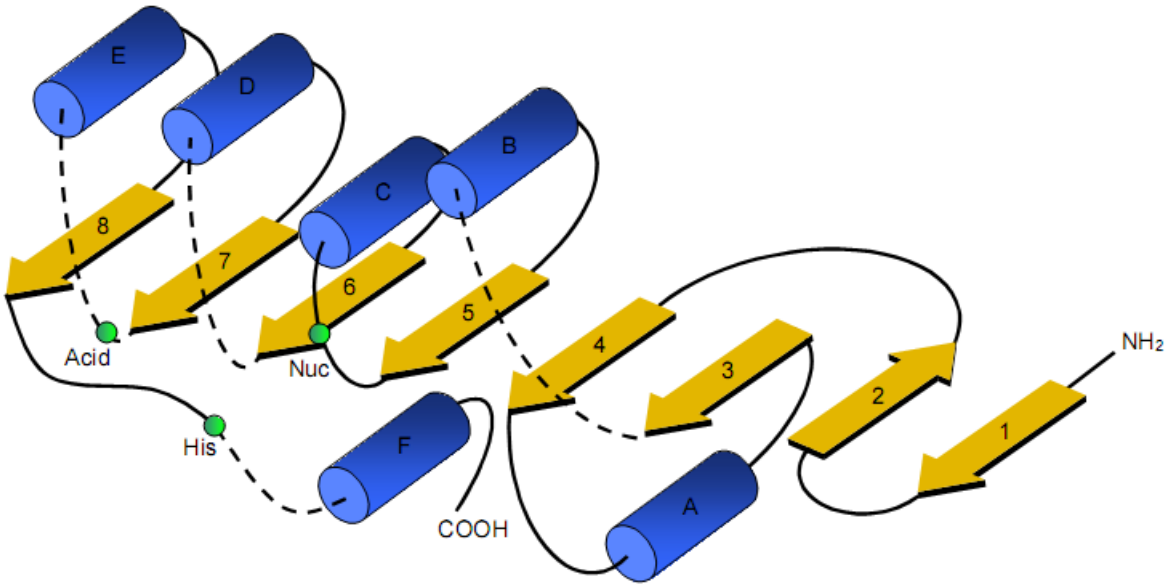


Abbildung 5-5: Schematische Darstellung des α/β-Hydrolase-*folds*. α-Helices sind als blaue Zylinder (A-F) dargestellt, β-Stränge als gelbe Pfeile (1-8). Die gestrichelten Linien deuten variable Bereiche an. Die Positionen der Aminosäuren der katalytischen Triade sind als grüne Punkte dargestellt (Fischer and Pleiss 2003).

5.5.2. Zugang zur Substratbindungstasche

Die Form und die Zugänglichkeit der Substratbindungstasche sind in Bezug auf Substratspezifität und Aktivität ein entscheidendes Merkmal aller Enzyme. Bei vielen Enzymen, darunter auch etlichen α/β-Hydrolasen, befindet sich das aktive Zentrum auf der Proteinoberfläche und ist damit gut zugänglich für das Substrat. Lipasen bilden in dieser Hinsicht oft eine Ausnahme. Sie besitzen eine Konformation, in der das aktive Zentrum für das Substrat zugänglich ist (offene Form), und eine Konformation, mit einem für das Substrat unzugänglichen aktiven Zentrum (geschlossene Form). Verantwortlich für dieses Phänomen ist das sogenannte *lid*, ein flexibler Teil der Struktur, der den Zugang zur Substratbindungstasche blockieren kann (Tejo, Salleh et al. 2004). Die Struktur des *lids* sowie dessen Position auf Sequenzebene kann sich dabei zwischen verschiedenen Enzymen unterscheiden. Beispielsweise wird das *lid* der Lipase aus *Candida rugosa* von einer α-Helix nahe dem N-Terminus des Proteins gebildet. Das *lid* der *human pancreatic* Lipase hingegen befindet sich nahe dem C-Terminus. Trotz dieser Unterschiede befinden sich beide *lids* strukturell an derselben Position im geschlossenen Zustand der Enzyme und blockieren den Zugang zur Substratbindungstasche (Tejo, Salleh et al. 2004).

5.5.3. Lipase Engineering Database

Die *Lipase Engineering Database* (LED) ist eine relationale Datenbank (Fischer and Pleiss 2003), die Sequenzen, Strukturen und Annotationen von Proteinen des α/β-Hydrolase-*folds* enthält. Die für Teile dieser Arbeit benutzte Version (Release 2) enthält 6668 Sequenzeinträge, die sich in 125 homologe Familien und 37 Superfamilien aufteilen. Sequenzen wurden basierend auf Sequenzähnlichkeit homologen Familien und Superfamilien zugeordnet. Zusätzlich enthält die LED Multisequenz-Alignments aller homologen Familien sowie diverse Annotationen (Position der katalytischen Triade, Schwefelbrücken, Signalpeptide). Die LED ist öffentlich zugänglich unter http://www.led.uni-stuttgart.de/.

5.6. Thiamindiphosphat-abhängige Enzyme

Die Beschreibung der ersten Thiamindiphosphat (ThDP)-abhängigen Enzyme geht zurück ins Jahr 1937 (Schellenberger 1998). Seit diesem Zeitpunkt hat sich die Anzahl an bekannten ThDP-abhängigen Enzyme stark vermehrt (Jordan 2003; Frank, Leeper et al. 2007). ThDP-abhängige Enzyme katalysieren eine Vielzahl an Reaktionen und gehören zu den Proteinfamilien der Decarboxylasen, Oxidoreduktasen, Transketolasen und α-Ketoglutarat-Dehydrogenasen. Besonders bedeutend ist dabei ihre Fähigkeit, neben der Bildung oder Spaltung von C-S-, C-O- oder C-N-Bindungen, C-C-Knüpfungen enantioselektiv bilden zu können. Industriell relevant ist dabei unter anderem die Möglichkeit aus zwei Aldehyden 2-Hydroxyketone zu bilden, welche Synthesebausteine für pharmakologisch aktive Verbindungen darstellen (Jordan 2003). Alle ThDP-abhängigen Enzyme sind dabei auf ihren Cofaktor ThDP angewiesen. ThDP ist ein Phosphatester des Thiamins (Vitamin B1) und stellt dessen biologisch aktive Form dar. Die Bildung von ThDP ist auf eine ständige Zufuhr von Vitamin B1 angewiesen, da dieses Vitamin vom menschlichen Körper nicht selbst gebildet werden kann. Ein Mangel an Vitamin B1 führt zu schweren Krankheitssymptomen, darunter das Korsakow Syndrom (Kopelman, Thomson et al. 2009) oder Beriberi, eine komplexe Vitaminmangelerkrankung. Trotz der Vielzahl an katalysierten Reaktionen liegt den ThDP-abhängigen Enzymen ein gemeinsames Reaktionsprinzip zugrunde. Dies besteht in der Bildung und Stabilisierung eines Intermediates, das als „aktivierter Aldehyd“ bezeichnet wird (Mosbacher, Mueller et al. 2005). Dabei erfolgt stets zuerst die Aktivierung von ThDP durch

Abspaltung eines Wasserstoffs vom C2-Atom des Thiazoliumrings. Dies geschieht durch die Bildung eines 1',4'-Imino-Tautomers des Pyrimidinrings. Dabei nimmt der Cofaktor eine V-Konformation ein. Diese ist essenziell, da dadurch das Proton des C2-Atoms des Thiazoliumrings in einen reaktionsfähigen Abstand zur 4'-Imino-Gruppe des benachbarten Pyrimidinrings gebracht wird (Abbildung 5-6) (Lindqvist, Schneider et al. 1992; Frank, Leeper et al. 2007). Das hochreaktive Intermediat kann daraufhin einen nukleophilen Angriff auf das jeweilige Substrat durchführen.

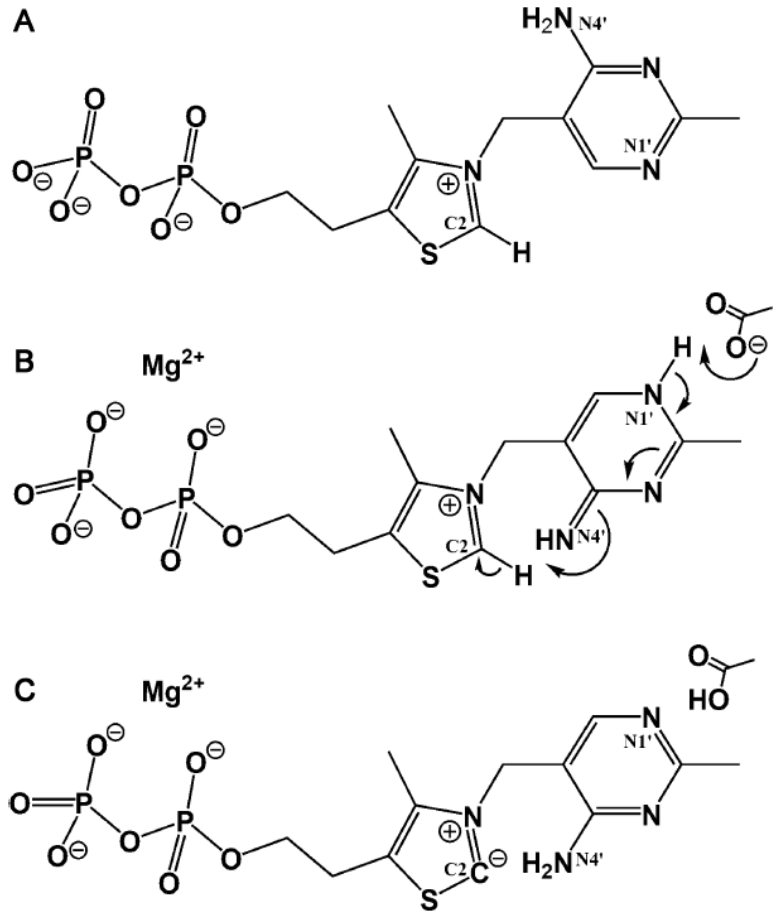


Abbildung 5-6: Chemische Struktur von ThDP. (A) bildet den Zustand in Lösung ab, (B) und (C) zeigen die aktivierte Form des Moleküls. (B) zeigt den Zwischenschritt der Aktivierung. ThDP ist an das Enzym gebunden und durch Mg^{2+} koordiniert. ThDP nimmt eine V-Konformation ein welche N4' in die Nähe von C2 bringt. Der von C2 abstrahierte Wasserstoff wird auf ein Glutamat des Enzyms übertragen. In (C) liegt das C2-Atom des Thiazoliumrings abstrahiert und damit aktiviert vor (Frank, Leeper et al. 2007).

5.6.1. Struktur der ThDP-abhängigen Enzyme

In der aktiven Form liegen die ThDP-abhängigen Enzyme als Di- oder Trimere vor. Die Proteine bestehen dabei immer aus mehreren Domänen, die jedoch, je nach Proteinfamilie, in Art und Anzahl abweichen können. Zwei Domänen sind jedoch bei allen Vertretern der ThDP-abhängigen Enzyme stets vorhanden: die Pyrimidin (PYR)-Bindedomäne und die Pyrophosphat (PP)-Bindedomäne. Diese sind essenziell für die Bindung und Aktivierung von ThDP. Die Domänen sind bei allen Vertretern der ThDP-abhängigen Enzyme strukturell sehr ähnlich, weisen jedoch starke Unterschiede in Bezug auf die Sequenz auf (Duggleby 2006). Die einzigen, in fast allen Proteinen konservierten Bereiche stellen dabei ein konserviertes, katalytisch aktives Glutamat innerhalb der PYR-Domäne und ein für die Bindung von Ionen (meistens Mg^{2+}) verantwortliches $GDX_{25-30}N$-Motiv innerhalb der PP-Domäne dar. Unter den weiteren, in wechselnder Reihenfolge und Anzahl vorhandenen Domänen befinden sich unter anderem die Transhydrogenase-dIII-Domäne (TH3) und die Transketolase C-terminale-Domäne (TKC) (Duggleby 2006; Costelloe, Ward et al. 2008) (Abbildung 5-7). Die Funktion der zusätzlich auftretenden Domänen ist in vielen Fällen nicht eindeutig geklärt und ihre Bedeutung in katalytischen Prozessen oft unklar (Costelloe, Ward et al. 2008).

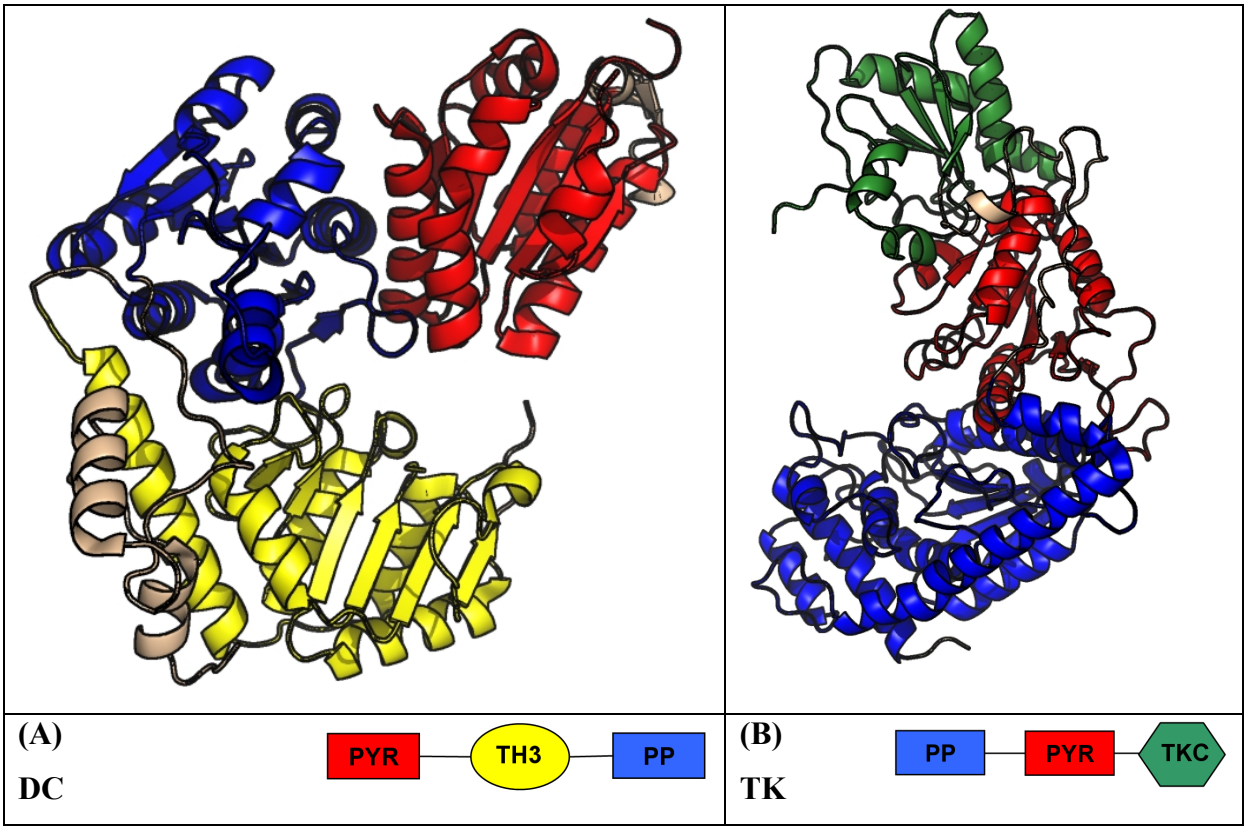


Abbildung 5-7: Schematische Darstellung der Domänen der DC- (Decarboxylase) und TK- (Transketolase) Familie der ThDP-abhängigen Enzyme. Die konservierten PP- und PYR-Domänen sind Bestandteil beider Proteine und strukturell sehr ähnlich, unterscheiden sich jedoch in ihrer Anordnung. (A) Proteine der DC-Familie enthalten eine TH3-Domäne, die sich zwischen der PYR- und PP-Domäne befindet. (B) Proteine der TK-Familie enthalten eine TKC-Domäne, die C-terminal vorliegt.

6. Zielsetzung

Im Rahmen dieser Arbeit wurden mehrere Proteinfamilien mit Hilfe von bioinformatischen Methoden systematisch analysiert sowie neue Proteinfamiliendatenbanken erstellt. Zu den betrachteten Proteinfamilien zählen die Familien der α/β–Hydrolasen, der Dehydrogenasen und Reduktasen sowie der ThDP-abhängigen Enzyme. Die Analysen sollten zu einem besseren Verständnis des Zusammenhangs zwischen Sequenz, Struktur und Funktion führen. Außerdem sollten sie zur Erstellung von allgemeingültigen Modellen zur Vorhersage des Expressions- und Aufreinigungsverhaltens dienen. Voraussetzung für diese Analysen war ein einheitliches Format der auszuwertenden Daten. Daher sollten Proteinfamiliendatenbanken erstellt beziehungsweise bestehende Proteinfamiliendatenbanken erweitert und aktualisiert werden. Je nach Proteinfamilie war es dabei notwendig, die Datenbank auf die auszuwertenden Informationen individuell anzupassen beziehungsweise die für die Analyse benötigten Informationen zu integrieren.

Anhand der Familie der α/β–Hydrolasen sollte ein Modell zur Vorhersage von funktionell relevanten seltenen Codons entwickelt werden. Das Ziel war es, mittels vergleichender Sequenzanalysen Codons zu identifizieren, die über ihre niedrige Nutzungsfrequenz Einfluss auf die Expression der jeweiligen Proteine nehmen. Durch Sequenz-Strukturvergleiche sollten die Positionen dieser Codons auf ihre strukturelle Lokalisierung und ihre Verteilung über die Proteinstruktur hin analysiert werden.

Anhand mehrerer Proteinfamiliendatenbanken sollten experimentell beobachtete Phänomene bei der Aufreinigung von Proteinen mittels Ionenaustauschchromatographie erklärt werden. Die aufgrund des isoelektrischen Punktes erwarteten pH-Werte für Bindung und Elution wichen teilweise stark von gemessenen Werten ab. Dieses abweichende Verhalten wurde verschiedenen Faktoren zugeordnet, darunter dem pI und der Anzahl der Histidine. Der Einfluss dieser Faktoren sollte modelliert und durch systematische Analysen ein möglicher Zusammenhang zwischen den Faktoren und dem Verhalten der Proteine identifiziert werden.

Die zum Zeitpunkt dieser Arbeit bereits bestehende Proteinfamiliendatenbank der Familie der α/β-Hydrolasen, die *Lipase Engineering Database* (LED) sollte erweitert und aktualisiert werden. Im Rahmen dieser Arbeit sollte zudem die neue Proteinfamilie der *Candida*

antarctica Lipase A in das bestehende Datenmodell integriert werden. Zudem sollten Hinweise zu dem bisher unbekannten Substratbindemechanismus der Lipase A durch Strukturvergleiche ermittelt werden.

Für die Familie der ThDP-abhängigen Enzyme sollte eine familienspezifische Datenbank erstellt werden. Zum besseren Verständnis dieser Proteinfamilie sollten die strukturell konservierten Bereiche jeder Familie identifiziert und annotiert werden. Die Datenbank sollte den familienübergreifenden Zugriff auf die Sequenzen dieser Bereiche ermöglichen und der systematischen Analyse familienspezifischer Eigenschaften dienen.

7. Ergebnisse

7.1. Analyse der Verteilung funktionell relevanter seltener Codons

(*siehe: Analysis of the distribution of functionally relevant rare codons, Seite 53*)

In der vorliegenden Arbeit wurde ein Modell zur Bestimmung funktionell relevanter seltener Codons entwickelt. Es beruht auf der Erkenntnis, dass seltene Codons zu einer Verlangsamung der Translationsrate an den Ribosomen führen können (Garel, Chavancy et al. 1981; Varenne, Buc et al. 1984; Curran and Yarus 1989; Mattes 1993; Thanaraj and Argos 1996; Musto, Romero et al. 2003). Die Geschwindigkeit der Translation wiederum hat Einfluss auf die Faltung des entstehenden Proteins. Eine korrekte Faltung eines Proteins ist oft nur möglich, wenn bestimmte Pausen während der Translation eingehalten werden. Diese geben der Peptidkette Zeit, die richtige Konformation einzunehmen. Werden Codons ausgetauscht und damit die Translationsgeschwindigkeiten verändert, kann dies zu Fehlfaltungen der Proteine führen (Thanaraj and Argos 1996; Li, Luo et al. 2003). Da seltene Codons zu Pausen während des Translationsprozesses führen, liegt der Schluss nahe, dass ein Austausch von seltenen Codons durch häufiger benutzte Codons dazu führen kann, dass auch solche Codons entfernt werden, die zur korrekten Faltung des Proteins beigetragen haben. Sollte die Verlangsamung der Translation ein Mechanismus des jeweiligen Organismus sein, um die korrekte Faltung des entsprechenden Proteins sicherzustellen, so ist davon auszugehen, dass dieser Mechanismus konserviert ist und im Laufe der Evolution erhalten blieb. Dies ist jedoch nicht gleichzusetzen mit der Konservierung der jeweiligen Aminosäure, da es für die meisten Aminosäuren sowohl häufig als auch weniger häufig genutzte Codons gibt. Hierdurch entstehen mehrere Möglichkeiten, eine bestimmte Aminosäure in Bezug auf ihre Translationsgeschwindigkeit zu codieren. Sollten seltene Codons eine essenzielle Funktion in der Expression bzw. der Faltung eines Proteins haben, so wäre zu erwarten, dass diese bei den Mitgliedern einer homologen Proteinfamilie, die denselben Faltungsweg beschreiten, konserviert sind. Da die Mitglieder einer homologen Proteinfamilie und eventuell sogar einer *fold*-Familie erwartungsgemäß einen ähnlichen Faltungsweg aufweisen (Kragelund, Hojrup et al. 1996; Clarke, Cota et al. 1999), sollte es einen evolutionären Druck auf die Konservierung von expressionsrelevanten Codons geben. Der Nachweis bzw. die Detektierung dieser seltenen Codons wurde daher in der vorliegenden Arbeit durch die Analyse von Multisequenz-Alignments homologer Proteinfamilien durchgeführt. Ein erster Ansatz wurde im Rahmen einer Diplomarbeit verfolgt (Widmann 2004) und beruhte auf der

vergleichenden Sequenzanalyse von Multisequenz-Alignments. Bei diesem Ansatz wurde das jeweilige Multisequenz-Alignment spaltenweise untersucht und zur Detektion seltener Codons ein Schwellenwert gewählt. Lag die Nutzungsfrequenz eines Codons unterhalb des Schwellenwerts wurde es als seltenes Codon gewertet. Die Analyse einzelner Spalten eines Multisequenz-Alignments hat jedoch den Nachteil, dass lediglich die Konservierung von Codons, die exakt an derselben Position im Alignment vorkommen, untersucht werden kann. Da jedoch ein kumulativer Einfluss von seltenen Codons in Bezug auf die Translationsgeschwindigkeit angenommen werden kann (Chou and Lakatos 2004), werden diese mit einem spaltenweise arbeitenden Modell nicht erkannt. Darüber hinaus besteht das Problem, ab welcher Frequenz ein Codon als selten zu definieren ist, da für diese Einteilung keine allgemeingültige Definition existiert. Zusätzlich muss bei Codons mit einer niedrigen Nutzungsfrequenz noch zwischen zwei Arten von Codons unterschieden werden: (1) Seltene Codons, die für eine Aminosäure codieren, die jedoch auch von anderen, häufiger genutzten Codons codiert werden kann (z. B. Arginin, für dessen Codierung sechs Codons zur Verfügung stehen). (2) Codons von Aminosäuren, die lediglich von ein bis zwei seltenen Codons codiert werden können (z. B. Tryptophan). Bei Codons der ersten Art kann davon ausgegangen werden, dass es für den Organismus möglich ist, die entsprechende Aminosäure durch ein seltenes oder ein häufig genutztes Codon zu codieren. Liegt nun eine Konservierung dieser seltenen Codons vor, kann davon ausgegangen werden, dass ein evolutionärer Druck besteht, der darauf abzielt an dieser Stelle einen langsam translatierenden Bereich sicherzustellen. Codons der zweiten Art sind hingegen stark von der jeweils codierten Aminosäure abhängig und wurden daher von der Analyse ausgeschlossen, da es keine Möglichkeit gibt, sie gegen häufiger genutzte Codons auszutauschen, ohne die codierte Aminosäure zu ändern. Ebenfalls wurde auf die Festlegung eines festen Schwellenwerts mangels einer allgemeingültigen Definition eines seltenen Codons verzichtet. Auf diesen Überlegungen basierend, wurde in der vorliegenden Arbeit anhand von 18 Proteinfamilien mit insgesamt 157 Proteinen ein von einem Schwellenwert unabhängiger Ansatz zur Identifizierung relevanter seltener Codons entwickelt. 16 Proteinfamilien wurden dabei der Proteinfamiliendatenbank der α/β–Hydrolasen, der *Lipase Engineering Database* (LED) entnommen. Zwei weitere Proteinfamilien wurden manuell erstellt und beinhalteten Mitglieder der Chloramphenicol-Acetyltransferase Proteinfamilie, sowie der Proteinfamilie der Fettsäure-bindenden Proteine. Mitglieder dieser Proteinfamilien waren zuvor das Ziel experimenteller Arbeiten gewesen, die detailliert die Auswirkungen eines Austauschs seltener Codons untersuchten. Bei dem in der vorliegenden Arbeit entwickelten Ansatz wurden

Codons nicht einzeln über ihre Nutzungsfrequenz als selten oder häufig vorkommende Codons klassifiziert. Zu diesem Zweck wurde für jede Proteinfamilie ein Multisequenz-Alignment erstellt. Das Ziel war es, jeder Spalte des Multisequenz-Alignments einen Wert zuzuordnen, der angibt, wie hoch der Anteil an seltenen Codons dieser Spalte ist. Zu diesem Zweck wurden die Frequenzen der in dieser Spalte vorkommenden Codons bestimmt und multipliziert, um die Frequenz der jeweiligen Spalte zu erhalten. Anschließend wurden alle möglichen Kombinationen aus synonymen Codons für die jeweilige Spalte bestimmt und die zugehörigen Frequenzen berechnet. Durch den Vergleich der Spaltenfrequenz der tatsächlich vorkommenden Codons mit den Frequenzen aller hypothetisch möglichen Kombinationen aus synonymen Codons war es möglich, den Anteil an seltenen Codons in jeder Spalte zu bestimmen, ohne dabei einen Schwellenwert definieren zu müssen. Für die Analyse des Multisequenz-Alignments wurde ein Fenster von jeweils 9 Spalten simultan ausgewertet. Dies geschah aufgrund des bereits erwähnten kumulativen Einflusses von seltenen Codons auf die Translationsgeschwindigkeit (Chou and Lakatos 2004). Die Fenstergröße von 9 wurde aufgrund der Tatsache gewählt, dass bis zu 27 Nukleotide während der Translation an das Ribosom binden (Zhang, Goldman et al. 1994). Regionen mit einer erhöhten Anzahl an konservierten seltenen Codons wurden als seltene Codon reiche Region (*rare codon rich region*, RCRR) klassifiziert.

Insgesamt wurden 42 RCRR Bereiche bei den untersuchten Proteinfamilien identifiziert. Die experimentell identifizierten Bereiche mit funktionell relevanten seltenen Codons konnten dabei erfolgreich vorhergesagt werden. 34 der identifizierten RCRRs konnten auf eine 3D-Struktur eines Proteins aus der jeweiligen Familie übertragen werden. Dies geschah, um eventuelle Präferenzen bzw. Anhäufungen von RCRRs in bestimmten Strukturelementen zu identifizieren. Die experimentell identifizierten, funktionell relevanten seltenen Codons befanden sich im Falle des Fettsäure-bindenden Proteins in einer l*oop*-Region (Cortazzo, Cervenansky et al. 2002), während sie im Falle der Chloramphenicol-Acetyltransferase sowohl in einem *loop* als auch in einem β-Strang zu finden waren (Komar, Lesnik et al. 1999). Diese Befunde demonstrierten, dass funktionell relevante seltene Codons sowohl in *loops* als auch in Sekundärstrukturelementen vorkommen können. Dies zeigte sich ebenfalls bei der hier durchgeführten Analyse. Sowohl in den Familien der beiden experimentell untersuchten Proteine als auch in den 16 Familien der α/β–Hydrolasen zeigte sich eine gleichmäßige Verteilung der konservierten Regionen auf *loop*-Regionen und Sekundärstrukturelemente. In beiden experimentell untersuchten Proteinen befanden sich sowohl die vorhergesagten RCRRs, als auch die experimentell identifizierten Bereiche in

Strukturelementen, die Proteindomänen verbinden. Im Fall des Fettsäure-bindenden Proteins befand sich das verbindende Element zwischen den beiden Hälften des β-*barrels*. Im Fall der Chloramphenicol-Acetyltransferase lag der Verbindungsbereich zwischen dem α– und dem β–*layer*. Aufgrund dieser Befunde liegt der Schluss nahe, RCRRs in Verbindung mit Strukturen zu bringen, die verschiedene Faltungsdomänen verbinden. Jedoch lieferte die Analyse der 16 Familien der α/β–Hydrolasen einen komplexeren Befund. Obwohl alle untersuchten Familien demselben *fold* angehören und daher erwartungsgemäß ein ähnliches Faltungsmuster aufweisen, verteilten sich die gefundenen RCRRs gleichmäßig über den repräsentativen α/β–Hydrolase-*fold*. Betrachtet man die Gesamtheit aller Fundstellen von RCRRs konnte kein hoch konservierter Bereich an RCRRs identifiziert werden, der von allen Mitgliedern der α/β–Hydrolasen geteilt wurde. Die höchste Konzentration an RCRRs aus verschiedenen Familien befand sich im Bereich der Helix D und umfasste insgesamt vier verschiedene Familien. Dabei ist jedoch zu beachten, dass die Region um Helix D einen hochvariablen Bereich innerhalb des sonst sehr gut konservierten α/β-Hydrolase-*folds* darstellt und je nach Familie einen wechselnden Anteil an Helices und *loop*-Bereichen enthält. Aufgrund dieser Erkenntnisse kann nicht davon ausgegangen werden, dass es einen für alle Mitglieder der α/β–Hydrolasen hoch konservierten Bereich an RCRRs gibt. Zudem wurden für 50 Prozent der untersuchten Proteinfamilien keine RCRRs gefunden. Für diese Beobachtungen gibt es drei mögliche Erklärungen: (1) Es gibt keine strukturell konservierten RCRRs, die essenziell für die Faltung aller α/β–Hydrolasen sind. Jedoch gibt es konservierte RCRRs innerhalb einzelner Proteinfamilien. (2) Die Mitglieder der α/β-Hydrolasen weisen keinen gemeinsamen Faltungsweg auf. Obwohl es Hinweise darauf gibt, dass Proteine mit einem einheitlichen *fold* auch einen gemeinsamen Faltungsweg beschreiten (Kragelund, Hojrup et al. 1996; Clarke, Cota et al. 1999), muss einschränkend erwähnt werden, dass diese Befunde aus Analysen vergleichsweise weniger Proteine stammen und daher eventuell nicht generalisiert werden können. (3) Der Grad an translationaler Selektion unterscheidet sich stark zwischen den verschiedenen Organismen der α/β–Hydrolase-Familie. In den meisten Organismen enthalten stark exprimierte Gene einen höheren Anteil an häufig genutzten Codons. Von dieser Regel weichen jedoch einige Organismen ab (dos Reis, Savva et al. 2004; Sharp, Bailes et al. 2005). Wie experimentell gezeigt wurde, kann ein Austausch seltener Codons in einer RCRR negative Folgen in Bezug auf die Expression des Proteins haben. Dies zeigt, dass die Analyse und Identifizierung von RCRRs für die Vorhersage von funktionell relevanten seltenen Codons genutzt werden kann. Der Austausch dieser Art von seltenen Codons im Rahmen einer Optimierungsstrategie sollte daher ausgeschlossen werden.

7.2. Die isoelektrische Region von Proteinen : Eine systematische Analyse

(siehe: The isoelectric region of proteins: a systematic analysis , Seite 79)

Die Ionenaustauschchromatographie stellt ein weit verbreitetes und etabliertes Verfahren zur Aufreinigung von Proteinen dar (Palekar, Vasudevan et al. 2000; Ahamed, Ottens et al. 2006). Die Bindung der Proteine an die Säule wird dabei größtenteils durch elektrostatische Wechselwirkungen zwischen den Proteinen und der geladenen Oberfläche der Säule bestimmt (Sheehan and FitzGerald 1996; Hallgren, Kalman et al. 2000). Aus diesem Grund hat sich die Vorhersage des isoelektrischen Punktes (pI) für eine erste Einschätzung des benötigten Auftragungs- bzw. Elutions pH-Werts bewährt (Ahamed, Nfor et al. 2007). Auftragungs pH-Werte werden dabei meist so gewählt das sie 0.5 – 1 pH-Einheit unter oder über dem pI des jeweiligen Proteins liegen (Healthcare 2004; Ahamed, Chilamkurthi et al. 2008). Es konnte jedoch gezeigt werden, dass viele Proteine bei pH-Werten binden bzw. eluieren, die mit dem vorhergesagten bzw. gemessenen pI nicht erklärbar sind (Ahamed, Nfor et al. 2007; Trodler, Nieveler et al. 2008). Eine detaillierte Analyse der pH-Werte, bei denen an eine Anionenaustauschersäule gebundene Proteine eluierten, wurde mittels eines pH-Gradienten durchgeführt (Ahamed, Nfor et al. 2007). Dabei zeigte sich, dass Proteine mit einem pI zwischen 6 und 8 bei pH-Werten eluierten, die deutlich über dem pI des jeweiligen Proteins lagen. Proteine mit einem pI kleiner als 6 oder größer als 8 hingegen eluierten an pH-Werten, die ihrem pI entsprachen (Ahamed, Nfor et al. 2007). Dieses Verhalten von Proteinen mit einem pI zwischen 6 und 8 wurde durch die Beobachtung erklärt, dass ihre Titrationskurven einen nahezu ladungsfreien Bereich in der Nähe des pIs aufwiesen, der sich teilweise über mehrere pH-Einheiten erstreckte (Ahamed, Nfor et al. 2007). Dieser verlängerte den ladungsneutralen Bereich des Proteins und führte dazu, dass die Proteine früher als erwartet ihre Ladung verloren und dadurch eluierten. In der vorliegenden Arbeit wurde für diesen Bereich der Begriff der *isoelectric region* (IER) eingeführt. Ein anderes experimentelles Beispiel behandelte die Schwierigkeiten der Bindung eines Proteins (*Candida antarctica* Lipase B, CALB) an eine Kationenaustauschersäule. Obwohl das Protein einen pI von 6 aufwies, war es nicht möglich, oberhalb eines pH-Werts von 3 eine Bindung an die Säule zu erreichen (Trodler, Nieveler et al. 2008). Dieser Befund konnte durch theoretische Untersuchungen des Proteins erklärt werden. CALB weist eine extrem große IER auf und verfügt damit über einen verlängerten Bereich, in dem das Protein nicht bzw. kaum geladen ist. Dadurch bindet es nicht wie erwartet in der Nähe des pIs an eine Kationenaustauschersäule. Zusätzlich wurde in dieser Arbeit die Vermutung aufgestellt, dass

die Größe der IER maßgeblich von der Anzahl an Histidinen des Proteins bestimmt wird. Dies wurde aus der Tatsache geschlossen, dass Histidin die einzige titrierbare Aminosäure im pH-Bereich von 5 bis 9 ist und somit maßgeblich den Verlauf der Titrationskurve in diesem Bereich bestimmt. Um festzustellen, ob und gegebenenfalls welchen Einfluss der pI bzw. die Anzahl der Histidine auf die IER haben, wurde eine systematische Analyse von zwei Proteinfamilien, der Familie der α/β–Hydrolasen mit über 4600 Sequenzen und der Dehydrogenase/Reduktase-Familie mit über 2600 Sequenzen durchgeführt. Zu beiden Proteinfamilien lagen bereits Proteinfamiliendatenbanken vor. Für die Proteinfamilie der α/β–Hydrolasen war dies die *Lipase Engineering Database* (LED) (Fischer and Pleiss 2003), für die Dehydrogenase/Reduktase-Familie die *Medium-Chain Dehydrogenase/Reductase Engineering Database* (Knoll and Pleiss 2008). Diese Ergebnisse wurden der Analyse eines Sets aus 5000 Zufallssequenzen gegenübergestellt. Es konnte gezeigt werden, dass Proteine mit einer großen IER nur einen Bruchteil aller Proteine darstellten, der unabhängig von der untersuchten Proteinfamilie bei ca. 2% lag. Die Untersuchung der vermuteten Faktoren mit maßgeblichem Einfluss auf die Größe der IER zeigte, dass weder der pI noch die Anzahl der Histidine allein mit der Größe der IER korrelierten. Erst die Kombination mehrerer Faktoren erlaubte die Unterteilung der Proteine in Gruppen, die eine klare Korrelation in Bezug auf die Anzahl der Histidine und die Größe der IER zeigten. Der dazu eingeführte Faktor B setzte sich aus dem Verhältnis der sauren zu basischen Aminosäuren und der Gesamtzahl an titrierbaren Aminosäuren zusammen. Dadurch ist es möglich, Proteine zu identifizieren, deren IER maßgeblich von der Anzahl an Histidinen bestimmt wird. Dies eröffnet darüber hinaus die Möglichkeit, gezielt die IER von Proteinen bei minimalen Änderungen an der Proteinsequenz zu beeinflussen. Daten über elektrostatische Eigenschaften der Proteine, wie der pI und die Größe der IER, wurden in das Datenmodell der LED integriert.

7.3. Strukturelle Einordnung mittels der *lipase engineering database*: Eine Fallstudie anhand der Lipase A aus *Candida antarctica*

(siehe: Structural classification by the Lipase Engineering Database: a case study of Candida antarctica lipase A , Seite 103)

Die *Lipase Engineering Database* (LED) integriert Sequenz- und Strukturinformationen von Lipasen, Esterasen und anderen verwandten Proteinen des α/β–Hydrolase-*folds*. Die LED ist öffentlich zugänglich unter der Adresse http://www.led.uni-stuttgart.de/. In der vorliegenden Arbeit wurde die neue Version 3.0 der LED erstellt und die Anzahl der enthaltenen Proteine stark erhöht. Im Vergleich zur vorhergehenden Version stieg die Anzahl der enthaltenen Proteineinträge von 4322 auf 18587 und die Anzahl der enthaltenen Strukturen von 167 auf 656. Neben einer reinen Aktualisierung der Datenbank und einer Erweiterung der bestehenden Proteinfamilien wurde eine neue Superfamilie eingeführt, die "*Candida antarctica lipase A like*" Superfamilie. Die Lipase A aus *Candida antarctica* (CALA) besitzt eine geringe Sequenzidentität zu anderen Lipasen und die zugehörige Kristallstruktur wurde erst kürzlich veröffentlicht (Ericsson, Kasrayan et al. 2008). Bei dieser Struktur handelt es sich um eine geschlossene Form der Lipase, bei der der Zugang zur Substratbindungstasche blockiert ist. Durch die nun zugängliche Kristallstruktur war es möglich, mittels vergleichender Strukturanalysen CALA im Rahmen der LED einzuordnen. Dazu wurde die ähnlichste Kristallstruktur innerhalb der LED identifiziert, die der Superfamilie der Deacetylasen angehört. Ein Vergleich beider Strukturen zeigte, dass trotz der strukturellen Ähnlichkeit entscheidende Unterschiede in Bezug auf die Struktur des *caps* und die Substratbindungstasche beider Enzymfamilien bestehen. Aufgrund dieser Unterschiede konnte die Struktur der Deacetylase nicht genutzt werden, um einen möglichen Zugang zur Substratbindungstasche der geschlossenen Form von CALA zu identifizieren. Bei weiteren vergleichenden Strukturanalysen wurde daher versucht, Proteinstrukturen mit einem ähnlichen Aufbau der Substratbindungstasche zu identifizieren. Wie gezeigt werden konnte, besitzt die Lipase aus *Candida rugosa* trotz einer geringeren Strukturähnlichkeit eine sehr ähnliche Substratbindungstasche (Abbildung 7-1). Durch einen strukturellen Vergleich konnte das mutmaßliche *lid* von CALA identifiziert werden, das bei beiden Strukturen eine sehr ähnliche Position einnimmt und den direkten Zugang zur Substratbindungstasche blockiert.

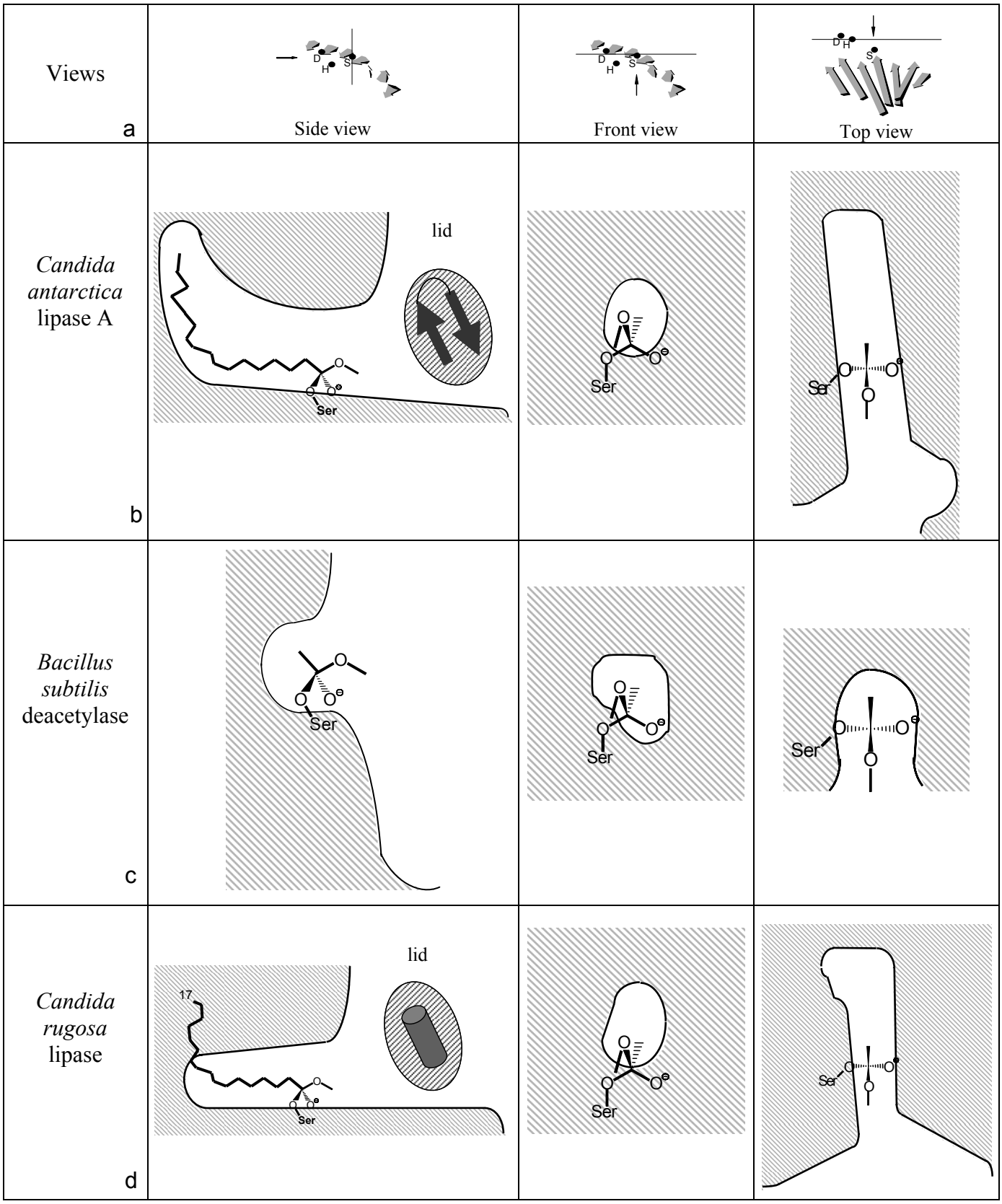


Abbildung 7-1: Form der Bindetaschen von *Candida antarctica* Lipase A (2VEO), *Bacillus subtilis* Deacetylase (1L7A) und *Candida rugosa* Lipase (1CRL).

7.4. Die Datenbank Thiamindiphosphat-abhängiger Enzyme: Untersuchung von Sequenz und Strukturbeziehungen

(*siehe: The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations, Seite 123)*

Thiamindiphosphat (ThDP)-abhängige Enzyme stellen eine große und vielfältige Enzymfamilie dar, die eine Vielzahl an enzymatischen Reaktionen katalysiert. Besonders bedeutend ist dabei ihre Fähigkeit, neben der Bildung oder Spaltung von C-S-, C-O- oder C-N-Bindungen, C-C-Knüpfungen enantioselektiv bilden zu können. Obwohl die verschiedenen Mitglieder der ThDP-abhängigen Enzyme höchst unterschiedlich in Bezug auf die Sequenz und die Anordnung der einzelnen Proteindomänen sind, teilen sie zwei gemeinsame Merkmale (Abbildung 7-2). Die Pyrophosphat (PP)- und Pyrimidin (PYR)-Domäne sind bei allen Mitgliedern strukturell konserviert, besitzen jedoch auf Sequenzebene nur eine sehr geringe Ähnlichkeit zwischen den verschiedenen Familien. Um eine umfassende und systematische Einteilung und Analyse der verschiedenen Mitglieder der ThDP-abhängigen Enzyme zu ermöglichen, wurde im Rahmen dieser Arbeit die *(ThDP)-dependent Enzyme Engineering Database* (TEED) entwickelt. Die TEED enthält derzeit 12048 Sequenzeinträge, die 9443 unterschiedlichen Proteinen und 379 Struktureinträgen zugeordnet sind, und kann über die Adresse www.teed.uni-stuttgart erreicht werden. Die Proteine wurden in insgesamt 8 Superfamilien und 64 homologe Familien unterteilt. Zu jeder Familie der TEED stehen Multisequenz-Alignments, phylogenetische Bäume sowie familienspezifische *Hidden Markov* Modelle (HMM) zur Verfügung. Die je nach Familie unterschiedliche Anordnung der Pyrophosphat- und Pyrimidin-Domänen machte es notwendig, diese Domänen für jede Proteinfamilie einzeln zu identifizieren. Beide Domänen wurden für jede Proteinfamilie individuell annotiert. Dies ermöglicht den direkten Zugriff auf die Sequenzen beider Domänen und macht systematische Analysen und Vergleiche auch über Superfamiliengrenzen hinweg möglich. ThDP-abhängige Enzyme humanen Ursprungs sind dafür bekannt, bei vielen Krankheiten, darunter Alzheimer und Diabetes, eine entscheidende Rolle zu spielen. Bei vielen Krankheitsbildern spielt bereits der Austausch einzelner Aminosäuren (*single nucleotide polymorphisms,* SNPs) innerhalb der Enzyme eine wichtige Rolle. Aufgrund der medizinischen Bedeutung der humanen ThDP-abhängigen Enzyme wurden diese einer systematischen Analyse unterzogen. Insgesamt enthält die TEED 66 Sequenzeinträge humanen Ursprungs, darunter mehrere Isoformen. Diese Sequenzeinträge wurden insgesamt

20 verschiedenen Proteinen zugeordnet. Mehr als 20 Sequenzen humanen Ursprungs, die in der *GenBank* entweder als unbekanntes oder mutmaßliches Protein annotiert sind, konnten durch die Analyse eindeutig einem bekannten Protein oder einer Proteinfamilie zugeordnet werden.

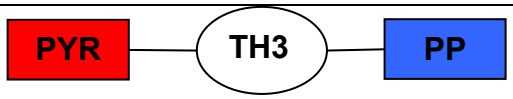
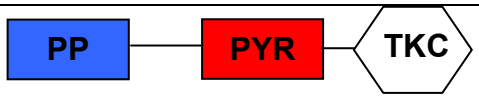
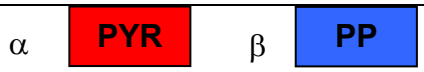
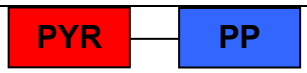
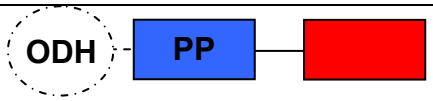
Superfamilie ID	Superfamilie	Strukturelle Anordnung der Proteindomänen
1	DC	
2	TK	
3	OR	
4	K1	
5	K2	
6	SPDC	
7	PPDC	
8	KDH	

Abbildung 7-2: Strukturelle Anordnung der Proteindomänen der 8 Superfamilien der TEED. Alle Proteinfamilien sind mit ihrer Superfamilie ID, dem Superfamiliennamen und einem Schema der strukturellen Anordnung der Domänen abgebildet.

8. Publikationen in englischer Sprache

1. Analysis of the distribution of functionally relevant rare codons
2. The isoelectric region of proteins: a systematic analysis
3. Structural classification by the Lipase Engineering Database: a case study of Candida antarctica lipase A
4. The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations

8.1. Analysis of the distribution of functionally relevant rare codons

Michael Widmann,[1] Marie Clairo,[2] Jürgen Dippon,[2] and Jürgen Pleiss [1§]

[1] Institute of Technical Biochemistry, Allmandring 31, 70569 Stuttgart

[2] Institut für Stochastik und Anwendungen, Pfaffenwaldring 57, 70569 Stuttgart

Publikation erschienen bei *BMC Genomics* 2008, 9: 207

8.1.1. Abstract

Background

The substitution of rare codons with more frequent codons is a commonly applied method in heterologous gene expression to increase protein yields. However, in some cases these substitutions lead to a decrease of protein solubility or activity. To predict these functionally relevant rare codons, a method was developed which is based on an analysis of multisequence alignments of homologous protein families.

Results

The method successfully predicts functionally relevant codons in fatty acid binding protein and chloramphenicol acetyltransferase which had been experimentally determined. However, the analysis of 16 homologous protein families belonging to the α/β hydrolase fold showed that functionally rare codons share no common location in respect to the tertiary and secondary structure.

Conclusion

A systematic analysis of multisequence alignments of homologous protein families can be used to predict rare codons with a potential impact on protein expression. Our analysis showed that most genes contain at least one putative rare codon rich region. Rare codons located near to those regions should be excluded in an approach of improving protein expression by an exchange of rare codons by more frequent codons.

8.1.2. Background

The usage of codons is not random and differs between organisms and genes. Depending on the strength of an organism's translational selection, there is a bias in highly expressed genes to avoid rare codons because of the low concentration of the respective tRNA in the cell (Ikemura 1981) which results in a decrease of translation rates (Varenne, Buc et al. 1984). As a consequence, genes with a high percentage of rare codons generally are translated at a lower rate than genes with a low percentage of rare codons (Pedersen 1984). Therefore, in an effort to increase the yield of recombinant proteins, rare codons have been replaced by more frequently used codons which led to increased yields of active protein (Makoff, Oxer et al. 1989; Zhou, Schnake et al. 2004).

However, gene redesign can also lead to abnormal protein folding and thus a decrease in protein solubility (Cortazzo, Cervenansky et al. 2002) as well as a decrease in protein activity (Crombie, Swaffield et al. 1992; Komar, Lesnik et al. 1999). It has been suggested that the differences in translational speed and the occurrence of pauses in translation is tightly linked to the folding mechanisms of the respective protein (Thanaraj and Argos 1996; Makhoul and Trifonov 2002), with clustered rare codons having a greater effect on translational speed than separated rare codons (Zhang, Goldman et al. 1994). Thus, optimal expression seems to be a consequence of a delicate balance between the occurrence and position of frequent and rare codons. Therefore, the effect of a replacement of rare by frequent codons to the expression level is not obvious. The goal of this work was to classify rare codons as critical and non-critical for expression of a given gene product. Non-critical rare codons could then be safely replaced by more frequent codons, while critical rare codons should not be replaced.

We suppose that critical rare codons can be predicted by comparing the codon usage of homologous proteins in a multisequence alignment. Therefore, we developed a new, cutoff independent approach to assign critical rare codons which compares the observed codon composition of one column in a multisequence alignment to all possible, alternative combinations of synonymous codons. Because the folding pathway of homologous proteins is assumed to be similar, rare codon rich regions (RCRR) which play a critical role in protein folding should be conserved in all members of a protein family. Since there is an increased probability to find rare codons in loop and linker regions (Thanaraj and Argos 1996), the location of RCRRs in respect to secondary structure elements was analyed.

This analysis was applied to two proteins for which it was experimentally shown that an exchange of rare codons with more frequent, synonymous codons reduces activity (Komar, Lesnik et al. 1999; Cortazzo, Cervenansky et al. 2002). The analysis of RCRRs was extended to systematically analyse a complete fold family. 16 protein families with a common α/β hydrolase fold were investigated to predict RCRRs, to localize them in respect to secondary and tertiary structure, and to identify possible RCRRs that are conserved in all members of the fold family.

8.1.3. Results

Fatty acid binding protein family

A protein family of homologues to fatty acid binding protein from *E. granulosus* consisting of 10 sequences was constructed and examined for rare codon rich regions (RCRRs). Sequence identities between the sequences ranged from 82% (fatty acid binding protein from *Taenia solium* as compared to *Echinococcus granulosus*) to 37% (*Taenia solium* / *Rattus norvegicus*). Two rare codon rich regions of 9 residues each were identified in the fatty acid binding protein family with scores of 1.8 and 2.6 respectively (Figure S1 in the supplementary file 1). Both RCRRs were mapped onto the 3D structure (Figure 1) of *E. granulosus* fatty acid binding protein (PDB: 1O8V). The fatty acid binding protein belongs to the β-barrel fold family. The barrel is formed by two antiparallel β–sheets: sheet 1 (β2-β5) and sheet 2 (β6-β10 and β1) are connected by an antiparallel pair of α-helices between β1 and β2 (Figure 2). The RCRRs are located at the connection between the two β–sheets: the first RCRR (G_{24}VDFVTRKM_{32}) comprises the loop connecting the two α-helices and the first turn of the second helix, the second RCRR (D_{77}SREVASLI_{85}) comprises the loop between strand β5 and β6 and 4 residues of the β6 strand. Previously it has been experimentally shown that the exchange of three rare codons by frequent synonymous codons in the region of the first RCRR ($R_{22}L_{23}G_{24}$) leads to misfolding as concluded from a significant drop in protein solubility and induction of stress response [6].

Chloramphenicol acetyltransferase protein family

A protein family of homologues to chloramphenicol acetyltransferase from *M. haemolytica* consisting of 8 sequences was constructed and examined for rare codon rich regions (RCRRs). Sequence identities between the sequences ranged from 82% (chloramphenicol acetyltransferase from *Yersinia pestis biovar* as compared to *Salmonella typhimurium*) to 34% (*Enterococcus faecium* / *Salmonella typhimurium*). Four rare codon rich regions with scores of 2.8, 3.6, 2.6 and 4.8 and lengths of 9, 11, 9 and 16 respectively were identified (Figure S2 in the supplementary file 2). The four RCRRs were projected on the 3D structure (Figure 3) of the *E. coli* chloramphenicol acetyltransferase (PDB: 1CIA). The chloramphenicol acetyltransferase protein belongs to the α/β class of proteins, forming a 2-layer sandwich consisting of a β-sheet and a layer of α-helices (Figure 4). The first RCRR is located in a loop region connecting two α-helices in the α–layer (S_{42}LDDSAYKF_{50}). The second RCRR is

located in a long loop region leading back to the β–layer and includes the major part of a β–strand (V_{79}WDSVDPQFTV_{89}). The third RCRR starts in a loop connecting the β–layer and the α–layer and includes a part of a helix of the α–layer (Y_{104}SSDIDQFM$_{112}$). The fourth RCRR consists of 16 amino acids and starts in the loop connecting this helix to the next β–strand of the β–layer, including this strand (K_{127}LFPQGVTPENHLNIS$_{142}$). Previously it has been experimentally shown that the exchange of a series of rare codons by frequent synonymous codons downstream of the third RCRR and overlapping with the fourth RCRR (S_{124}DTKLFPQGVTPENHLNISAL$_{144}$) supposedly led to the elimination of a translational pause in this region and caused a drop in specific activity by 20% (Komar, Lesnik et al. 1999).

α/β hydrolase families

A set of 16 homologous protein families belonging to the same α/β hydrolase fold family were systematically compared (Tab. 1). To find out whether critical rare codons are preferentially located in loop regions rather than in α-helices or β-strands, the location of RCRRs in respect to secondary structure elements was analysed. In addition, comparing the location of RCRRs in proteins with different sequence but identical fold allows to investigate whether RCRRs are conserved on the level of fold, supposing that all proteins of the same fold have a similar bottleneck in the folding pathway. Therefore, each family was examined and the RCRRs were mapped onto a crystal structure if available. 16 protein families with 7 or more proteins per family were retrieved from the Lipase Engineering Database (LED (Fischer and Pleiss 2003)) and analyzed for RCRRs. 2 protein families (abH17.01 and abH24.01) contained RCRRs but no family member with crystal structure. Therefore, the RCRRs could not be assigned to secondary structure elements. 3 families contained no RCRRs (abH09.02, abH30.01, abH31.02). 5 families only contained putative RCRRs in highly diverse regions (abH14.02, abH23.01, abH26.01, abH28.01, abH33.01). In 6 families a total of 32 RCRRs were detected and mapped to the respective crystal structure (Tab. 1). 29 RCRRs could be unambiguously assigned to one of four groups, depending on their location in secondary structure elements: (1) completely located in a loop region, (2) mainly located in a loop region (more than 50% of the RCRR in a loop region), (3) mainly located in an α-helix or a β-strand (more than 50% of the RCRR in a α-helix or a β-strand), and (4) completely located in a secondary structure element (Tab. 2). 3 RCRRs could not be assigned to a group due to missing structure information in the crystal structure. Of the 29 assigned RCRRs, 6, 8,

11, and 4 RCRRs belong to groups 1 to 4, respectively. Thus, no preference of RCRRs for loop regions was observed.

To identify RCRRs that are conserved across family borders, the 32 RCRRs were mapped on the representative α/β fold and are displayed according to their respective window score (Figure 5). Multiple RCRRs in one family in the same region were considered as only one hit. The RCRRs are distributed over 17 different positions in the representative α/β fold: 14 positions with RCRRs from only one family, 1 position with RCRRs from 2 different protein families, 1 position with RCRRs from 3 different families, and 1 position with RCRRs from 4 families. The position with RCRRs from 3 different families is located in the loop region between β-strand 3 and α-helix B. The position with RCRRs from 4 different families is located in the region of α-helix D. This region is highly variable among the protein families and often consists of more than one helix.

8.1.4. Discussion

Cutoff-independent and unbiased prediction of rare codon rich regions

In most genes an exchange of rare codons with synonymous, more frequent codons is neutral or even increases the yield of soluble protein (Makoff, Oxer et al. 1989; Rangwala, Finn et al. 1992; Slimko and Lester 2003). For some genes, however, it has been observed that such an exchange surprisingly leads to an increase of incorrectly folded proteins (Komar, Lesnik et al. 1999; Cortazzo, Cervenansky et al. 2002; Yadava and Ockenhouse 2003). Therefore, we based our investigation on the hypothesis that there might exist rare codons which have a regulatory function in translation and contribute to the correct folding pathway of a protein. Because the members of a homologous family and probably also of a fold family are expected to have a similar folding pathway, there should be an evolutionary bias towards the conservation of these critical rare codons. Because we only analyse synonymous codons, we restrict our analysis to the observed amino acid sequence. Thus, a possible effect to the expression level upon exchange of an amino acid is not considered by our analysis.

A rare codon is usually defined by a low usage frequency. Two types of rare codons have to be distinguished: (1) rare codons that code for an amino acid that is also encoded by more frequent codons (e.g. the arginine codon AGG) and (2) rare codons of amino acids (e.g. W,Y,H) that are encoded by only one or two rare codons. Our rare codon analysis identifies the first type of rare codons. While these rare codons are supposed to be the result of a significant evolutionary pressure towards using a rare codon instead of a frequent codon at the respective position, the second type of rare codons is strongly biased toward positions with highly conserved amino acids that are encoded exclusively by rare codons. For many organisms, codon usage tables are available (Nakamura, Gojobori et al. 2000). However, a generally applicable distinction between rare and frequent codons is not available and the result of the analysis would depend on the choice of an arbitrary cutoff value. Therefore, we have developed a cutoff-independent approach to assign rare codons by comparing the observed codon composition of one column to all possible, alternative combinations of synonymous codons. For each column a quantitative rare codon score is derived. Instead of single columns, a sliding window of 9 columns is evaluated, because up to 27 nucleotides are involved in binding to the ribosome during translation (Zhang, Goldman et al. 1994) and a cumulative effect of neighbouring rare codons has been expected (Chou and Lakatos 2004).

Location of rare codon rich regions

It has been suggested that there is an increased tendency for rare codons in loop and linker regions (Thanaraj and Argos 1996; Thanaraj and Argos 1996; Komar, Lesnik et al. 1999). For two proteins being examined for RCRRs, functionally relevant rare codons have been experimentally identified which led to a decrease of expressed active protein upon exchange by more frequent codons. Interestingly, in the gene coding for a fatty acid binding protein, the functionally relevant rare codons are located in a loop region [6], while in the second gene, the chloramphenicol acetyltransferase, the functionally relevant rare codons are located both in a loop and in a β-strand [8]. The observation of functionally relevant rare codons located in both loop and secondary structure regions is confirmed by our analysis of rare codon rich regions which predicts about 50 % of RCRRs in loop and secondary structure regions, both in our analysis of the two experimentally examined genes and of 16 α/β hydrolase families. However, because our prediction of RCRRs is restricted to regions with a sufficient conservation of amino acids, highly diverse regions are excluded from the analysis. Therefore, functionally relevant rare codons could not be predicted if they were located in highly variable loop regions.

In the two experimentally investigated genes, RCRRs were predicted in regions linking the two halves of the β-barrel in the fatty acid binding protein and the α and β layer in the chloramphenicol acetyltransferase. Thus it is tempting to associate RCRRs with regions that link two separate folding domains. However, our systematic analysis of 16 α/β hydrolase families provides a more complex picture. Although all families are of the same fold and thus are expected to have a similar folding pathway the RCRRs are nearly equally distributed in the representative α/β hydrolase fold.

This holds true even when a more stringent cutoff is applied and RCRRs close to the minimal score requirement are eliminated. Taking all RCRRs into account, only two areas with an increased density of RCRRs are found. The region encompassing helix D with 4 RCRRs from 6 different families and the loop region connecting β-strand 3 to helix B with 3 RCRRs from 6 different families. However, the region encompassing helix D is highly variable among the α/β hydrolase families and consists of a varying number of strands and helices. The loop region connecting β-strand 3 to helix B connects the first half of the β-sheet to the second half, consisting of 4 β-strands each. Thus, there seems to be no common region in which RCRRs are located in all α/β hydrolases. In addition, 50% of all α/β hydrolase families contain no RCRRs at all. This observation can be explained by either of three possibilities: (1) There are no rare codons which are structurally conserved in all α/β hydrolases and are

essential to control folding. However, RCRRs were found in individual homologous families. (2) α/β hydrolases do not have a common folding pathway. While there is evidence that proteins sharing the same fold also share a common folding pathway (Clarke, Cota et al. 1999) (Kragelund, Hojrup et al. 1996), this observation was based on a small set of proteins and therefore can not be generalized. Indeed, there are some studies showing that proteins sharing a common structure undergo a different folding pathway *in vitro* (Widmann and Christen 1995; Ropson, Yowler et al. 2000). (3) The level of translational selection might differ among species. In most organisms highly expressed genes seem to contain a higher percentage of frequently used codons, while in 30% no such codon bias was found (Sharp, Bailes et al. 2005) (dos Reis, Savva et al. 2004). However, this method averages over the whole gene and therefore does not take local conservation of rare codons into account.

As it has been shown experimentally that replacing rare codons by more frequent codons in proximity to a RCRR can lead to a decrease in protein expression, the analysis of RCRRs could be helpful in predicting those critical rare codons which are probably beneficial to expression and should not be a target for codon replacement.

However, it seems that a prediction of RCRRs has to be restricted to single homologous families

8.1.5. Conclusions

In most cases the substitution of rare codons with more frequent codons leads to increased protein yields in heterologous gene expression. To predict functionally relevant rare codons, multisequence alignments were analyzed to identify conserved rare codon rich regions. The prediction was validated by experimental data on silent mutations of two proteins. Therefore, we suggest that the approach of improving protein expression by an exchange of rare codons by more frequent codons should exclude rare codons located in highly conserved rare codon rich regions. A systematic analysis of 16 α/β hydrolase families predicts that most genes contain at least one putative rare codon rich region. They are however not restricted to loop regions but also occur in secondary structure elements. In addition, no preferred location of rare codon rich regions was found in respect to the common α/β hydrolase fold.

8.1.6. Methods

Protein families

Two proteins were analysed which show decreased activity upon replacement of rare by frequent codons: fatty acid binding protein from *Echinococcus granulosus* (Cortazzo, Cervenansky et al. 2002) and chloramphenicol acetyltransferase III from *Escherichia coli* (Komar, Lesnik et al. 1999).

The protein and DNA sequences of proteins homologous to fatty acid binding protein and chloramphenicol acetyltransferase III were retrieved from the GenBank by a BLAST search (Altschul, Gish et al. 1990) starting with GenBank entries GenBank:Q02970 and GenBank:NP_073222, respectively. Only proteins from different organisms and with a sequence identity between 35% and 80% were selected for the subsequent multisequence alignment.

Protein and DNA sequences of 16 protein families (Tab. 1) with 7 or more proteins per family were extracted from the Lipase Engineering Database (Fischer and Pleiss 2003). The family classification scheme of the Lipase Engineering Database was used which led to some families with overall sequence identities of only 20 %. For 14 families representative structures were available in the PDB. Families with more than 10 members were reduced in size by excluding proteins from the same organism if possible, else sequences with the lowest sequence identity were removed.

A multisequence alignment of the protein sequences of each protein family was constructed using ClustalW (Thompson, Higgins et al. 1994) with a Gonnet Matrix (Gonnet, Cohen et al. 1992) and a gap opening and extension penalties of 10 and 0.2, respectively. For each protein sequence, the DNA sequence was retrieved and codons were assigned to the respective amino acid in the multisequence alignment.

Scoring method

For each column of the multisequence alignment, a codon score S was evaluated. For every amino acid, the usage frequency of its codon was taken from the Codon Usage Database (Nakamura, Gojobori et al. 2000). These frequencies were multiplied, resulting in the column frequency α. Then all possible codon combinations were determined and their respective frequencies multiplied, resulting in codon frequencies β_i for each combination i (i=1,N). Each column frequency β_i was then compared to the column frequency α, and the number n of all β_i

≤ α was determined. The score S of each column was evaluated by normalizing the number n by the number of all possible codon combinations N: S = n/N.

Small values of S correspond to a high percentage of rare codons. Thus, five groups were defined: group 1 of highly conserved rare codons with $0 \leq S < 0.2$, group 2 of conserved rare codons with $0.2 \leq S < 0.4$, group 3 with ($0.4 \leq S < 0.6$), group 4 with ($0.6 \leq S < 0.8$) and group 5 with ($0.8 \leq S \leq 1$). The number of columns belonging to each group was counted for each protein family and the total sum for each column group was determined (Tab. S3 in the supplementary file 3). From the total sums, the probability of each column group as well as the ratio between the groups was determined. To predict rare codon rich regions (RCRRs), a window of nine columns was analyzed by counting the numbers S_1 and S_2 of all columns belonging to group 1 and 2, respectively. The number of columns of group 1 and group 2 correspond to 2.5% and 4.5%, respectively, of all columns and have a ratio of 1.7. A window score W was evaluated by a weighted sum of S_1 and S_2. Because group 2 columns were 1.7 fold more frequent than group 1 columns, they were weighted with a factor of 0.6:

$$\mathbf{W = S_1 + S_2 * 0.6}$$

Thus each column of group 1 inside the window contributes a score of 1, while a column of group 2 contributes a slightly smaller score of 0.6. Areas with a window score $W \geq 1.8$ are designated as a putative RCRR, beginning from the first contributing column to the last one (columns of group one or two). This score was chosen in order to avoid the detection of single columns from group 1 as a putative RCRR. Thus, a putative RCRR is predicted if at least 2 columns of group 1, 1 column of group 1 and 2 columns of group 2, or 3 columns of group 2 are found. For both cases, the probability of a random occurrence was estimated using a binominal distribution: the probability of finding 2 columns of group 1 in a window of 9 columns is 2%, and the probability of finding three or more columns of either group 1 or group 2 is 2%. Therefore, the probability of randomly finding a putative RCRR is 4%. Neighbouring RCRRs with a distance of less than 9 columns are merged. Thus, these merged RCRR will exceed the initial window length of 9 columns. Each of the putative RCRRs were evaluated for the quality of the local multisequence alignment by PLOTCON from the EMBOSS suite (Rice, Longden et al. 2000) with the EBLOSUM62 matrix. To be accepted as an RCRR the average PLOTCON score of a detected putative RCRR has to be at least 1.0. Thus, putative RCRRs that are located in highly variable regions were rejected.

8.1.7. Abbreviations

RCRR, rare codon rich region

8.1.8. Authors' contributions

MW carried out the analysis and drafted the manuscript, MC contributed in developing the algorithm, JD contributed to the statistical analysis, and JP supervised the study.
All authors read and approved the final manuscript.

8.1.9. Acknowledgements

We thank the Federal Ministry of Education and Research (PTJ 0313434D) for financial support.

8.1.10. References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Chou, T. and G. Lakatos (2004). "Clustered bottlenecks in mRNA translation and protein synthesis." Phys Rev Lett **93**(19): 198101.

Clarke, J., E. Cota, S. B. Fowler and S. J. Hamill (1999). "Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway." Structure **7**(9): 1145-53.

Cortazzo, P., C. Cervenansky, M. Marin, C. Reiss, R. Ehrlich, et al. (2002). "Silent mutations affect in vivo protein folding in Escherichia coli." Biochem Biophys Res Commun **293**(1): 537-41.

Crombie, T., J. C. Swaffield and A. J. Brown (1992). "Protein folding within the cell is influenced by controlled rates of polypeptide elongation." J Mol Biol **228**(1): 7-12.

dos Reis, M., R. Savva and L. Wernisch (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic Acids Res **32**(17): 5036-44.

Fischer, M. and J. Pleiss (2003). "The Lipase Engineering Database: a navigation and analysis tool for protein families." Nucleic Acids Res **31**(1): 319-21.

Gonnet, G. H., M. A. Cohen and S. A. Benner (1992). "Exhaustive matching of the entire protein sequence database." Science **256**(5062): 1443-5.

Ikemura, T. (1981). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system." J Mol Biol **151**(3): 389-409.

Komar, A. A., T. Lesnik and C. Reiss (1999). "Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation." FEBS Lett

462(3): 387-91.

Kragelund, B. B., P. Hojrup, M. S. Jensen, C. K. Schjerling, E. Juul, et al. (1996). "Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family." J Mol Biol **256**(1): 187-200.

Makhoul, C. H. and E. N. Trifonov (2002). "Distribution of rare triplets along mRNA and their relation to protein folding." J Biomol Struct Dyn **20**(3): 413-20.

Makoff, A. J., M. D. Oxer, M. A. Romanos, N. F. Fairweather and S. Ballantine (1989). "Expression of tetanus toxin fragment C in E. coli: high level expression by removing rare codons." Nucleic Acids Res **17**(24): 10191-202.

Nakamura, Y., T. Gojobori and T. Ikemura (2000). "Codon usage tabulated from international DNA sequence databases: status for the year 2000." Nucleic Acids Res **28**(1): 292.

Pedersen, S. (1984). "Escherichia coli ribosomes translate in vivo with variable rate." Embo J **3**(12): 2895-8.

Rangwala, S. H., R. F. Finn, C. E. Smith, S. A. Berberich, W. J. Salsgiver, et al. (1992). "High-level production of active HIV-1 protease in Escherichia coli." Gene **122**(2): 263-9.

Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Ropson, I. J., B. C. Yowler, P. M. Dalessio, L. Banaszak and J. Thompson (2000). "Properties and crystal structure of a beta-barrel folding mutant." Biophys J **78**(3): 1551-60.

Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden and R. E. Sockett (2005). "Variation in the strength of selected codon usage bias among bacteria." Nucleic Acids Res **33**(4): 1141-53.

Slimko, E. M. and H. A. Lester (2003). "Codon optimization of Caenorhabditis elegans GluCl ion channel genes for mammalian cells dramatically improves expression levels." J Neurosci Methods **124**(1): 75-81.

Thanaraj, T. A. and P. Argos (1996). "Protein secondary structural types are differentially coded on messenger RNA." Protein Sci **5**(10): 1973-83.

Thanaraj, T. A. and P. Argos (1996). "Ribosome-mediated translational pause and protein domain organization." Protein Sci **5**(8): 1594-612.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Varenne, S., J. Buc, R. Lloubes and C. Lazdunski (1984). "Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains." J Mol Biol **180**(3): 549-76.

Widmann, M. and P. Christen (1995). "Differential effects of molecular chaperones on refolding of homologous proteins." FEBS Lett **377**(3): 481-4.

Yadava, A. and C. F. Ockenhouse (2003). "Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems." Infect Immun **71**(9): 4961-9.

Zhang, S., E. Goldman and G. Zubay (1994). "Clustering of low usage codons and ribosome movement." J Theor Biol **170**(4): 339-54.

Zhou, Z., P. Schnake, L. Xiao and A. A. Lal (2004). "Enhanced expression of a recombinant malaria candidate vaccine in Escherichia coli by codon optimization." Protein Expr Purif **34**(1): 87-94

8.1.11. Figures

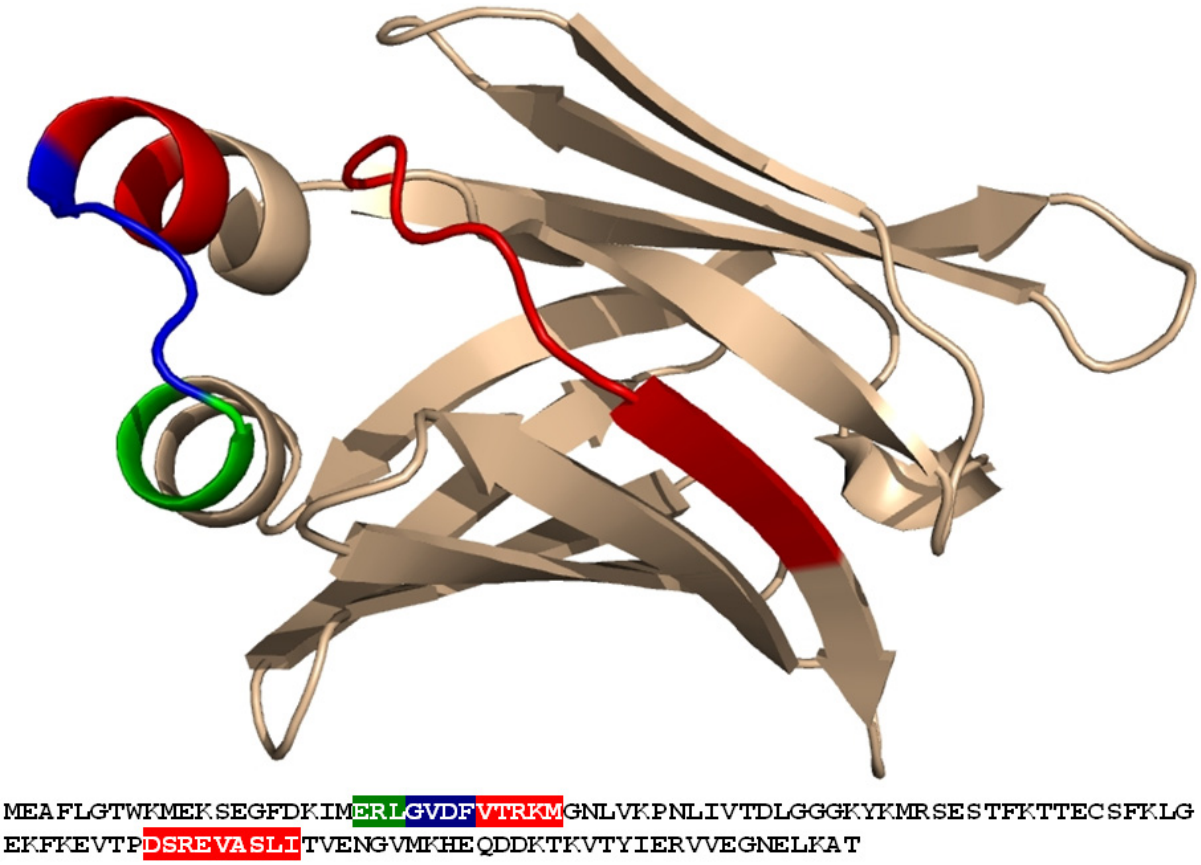


Figure 1 - Projection of rare codon rich regions on the sequence and the crystal structure (PDB entry 1O8V) of fatty acid binding protein. Regions containing RCRRs are colour coded in the sequence and the three dimensional structure: a region that contains the predicted RCRRs (red), the experimentally examined region (orange), a region that has been predicted and was also experimentally examined (blue).

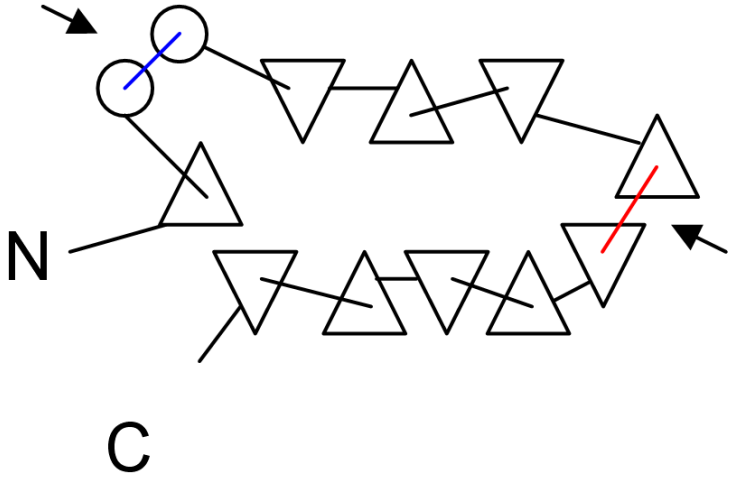


Figure 2 - 2D projection of the fatty acid binding protein 3D structure. View is from above towards the β-barrel. α-helices are represented as circles, β-strands as triangles. Upward and downward facing triangles represent β-strands directed upwards and downwards, respectively. Regions containing RCRRs are colour coded: a region that contains the predicted RCRRs (red) and a region that has been predicted and was also experimentally examined (blue).

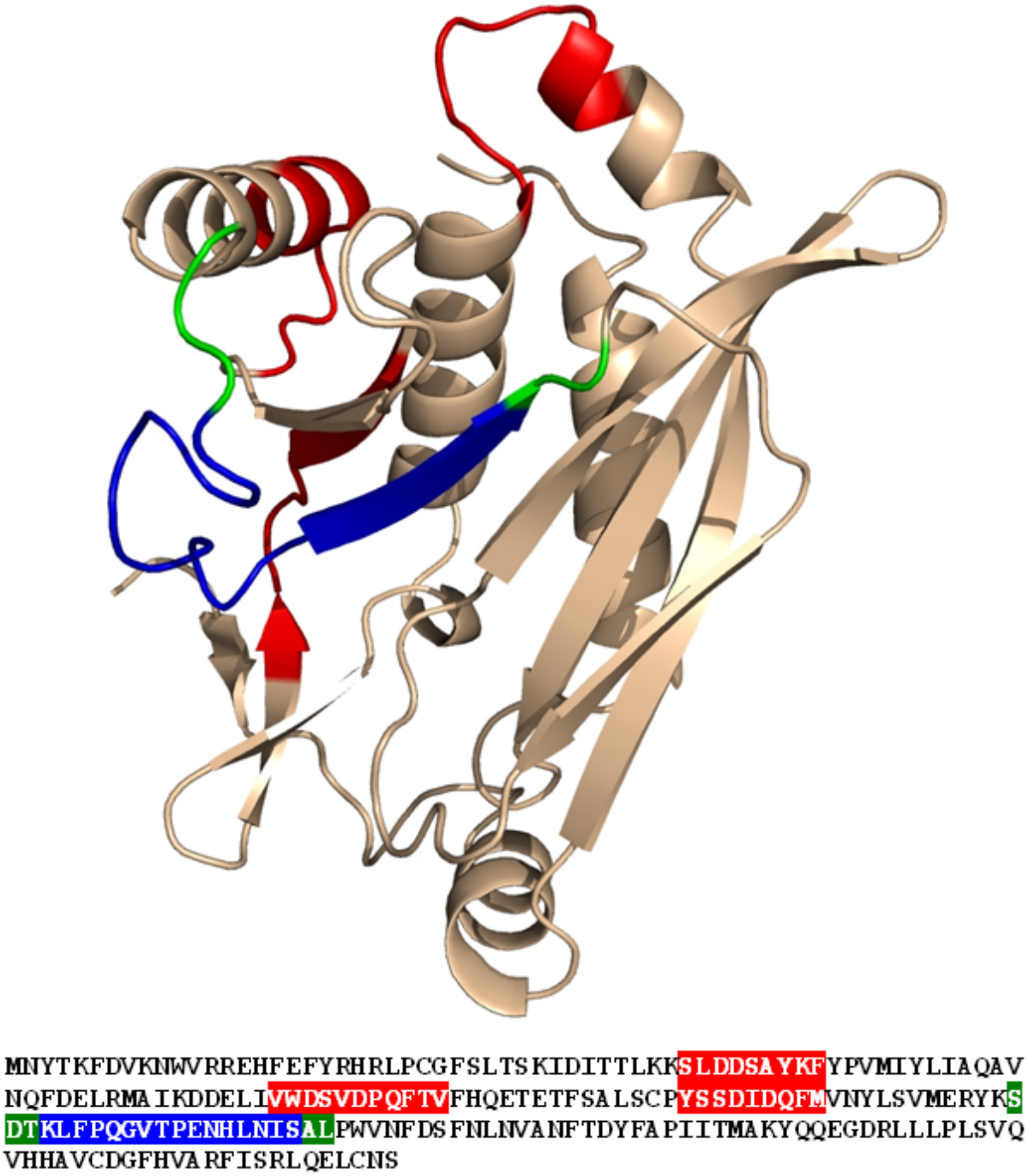


Figure 3 - Projection of rare codon rich regions on the sequence and the crystal structure (PDB entry 1CIA) of chloramphenicol acetyltransferase. Regions containing RCRRs are colour coded in the sequence and the three dimensional structure: a region that contains the predicted RCRRs (red), the experimentally examined region (orange), a region that has been predicted and was also experimentally examined (blue).

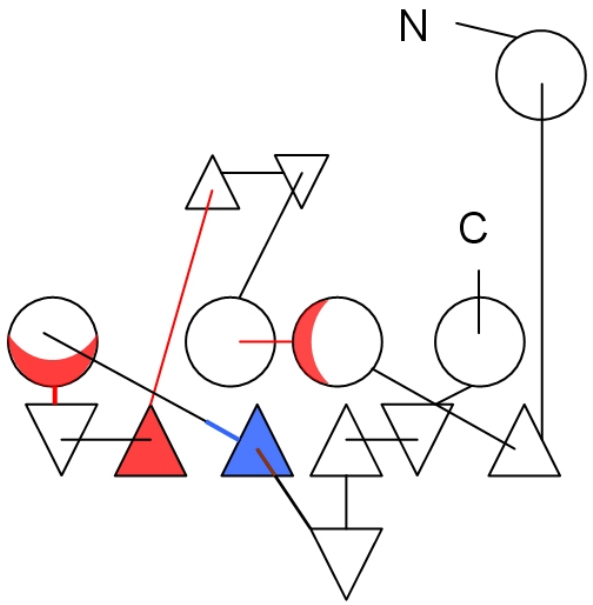


Figure 4 - 2D projection of the chloramphenicol acetyltransferase protein 3D structure. View is from above towards the β-barrel. α-helices are represented as circles, β-strands as triangles. Upward and downward facing triangles represent β-strands directed upwards and downwards, respectively. Regions containing RCRRs are colour coded: a region that contains the predicted RCRRs (red) and a region that has been predicted and was also experimentally examined (blue).

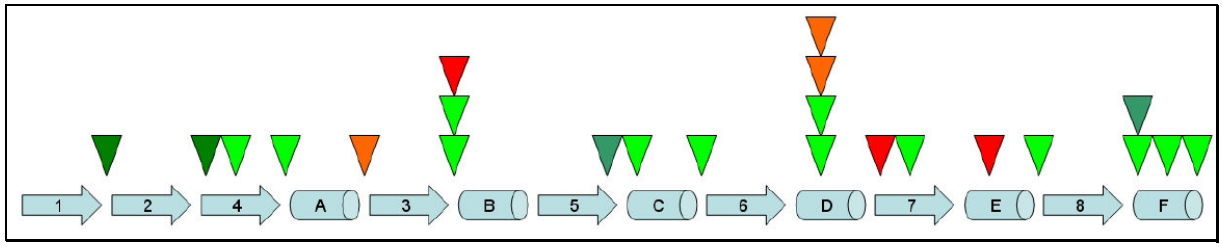


Figure 5 - Position and number of RCRRs, projected on a linear representation of the α/β fold. α–helices and β–strands are depicted as cylinders and arrows, respectively, the linking loops are not shown. Predicted RCRRs are represented by coloured triangles. Each triangle represents a RCRR in one distinct homologous protein family. Triangles are coloured by the respective window score W (light green $1.8 \leq W \leq 2.7$, dark green $2.8 \leq W \leq 3.7$, orange $3.8 \leq W \leq 4.7$, red $4.8 \leq W$).

8.1.12. Tables

Table 1 - Homologous protein families from the Lipase Engineering Database. All families are listed with their internal unique identifier (LED ID), their family name, the number of predicted RCRRs, the number of sequences in this family, and the PDB entry used to assign predicted RCRRs to secondary structure elements.

LED ID	Homologous family name	No. of RCRRs	No. of sequences	PDB-entry
abH01.02	*Mammalian carboxylesterases*	10	9	1K4Y
abH08.14	*Ccg1/TafII250-interacting factor B like*	2	9	1IMJ
abH09.02	*BioH protein like*	0	10	1M33
abH12.01	*Hydroxynitrile lyases*	3	10	1QJ4
abH14.02	*Gastric lipases*	0	10	1HLG
abH15.02	*Burkholderia cepacia lipase like*	6	7	4LIP
abH17.01	*Chloroflexus aurantiacus lipase like*	3	7	-
abH19.01	*Palmitoyl-protein thioesterase 1 like*	4	8	1EXW
abH23.01	*Rhizomucor mihei lipase like*	0	10	1DU4
abH24.01	*Pseudomonas lipases*	2	8	-
abH26.01	*Deacetylases*	0	7	1ODT
abH28	*Prolyl endopeptidases*	0	9	1O6F
abH30.01	*Cocaine esterases*	0	8	1L7Q
abH31.02	*Carboxymethylenebutenolidases*	0	8	1DIN
abH33.01	*Antigen 85-C*	0	10	1DQZ
abH34.02	*Serine carboxypeptidase II like*	7	9	1GXS

Table 2 - Number of predicted RCRRs in four groups of secondary structure elements. Groups: (1) completely located in a loop region, (2) mainly located in a loop region (more than 50% of the RCRR in a loop region), (3) mainly located in an α-helix or a β-strand (more than 50% of the RCRR in a α-helix or a β-strand), and (4) completely located in a secondary structure element. Families are referred by their internal database identifier (LED ID, see Table 1).

LED ID	Group			
	1	2	3	4
abH01.02	2	2	5	-
abH08.14	1	-	-	1
abH12.01	1	1	-	1
abH15.02	-	3	1	2
abH19.01	-	1	3	-
abH34.02	2	1	2	-

8.1.13. Supplementary material

```
Salmo_salar                 --MAEAFAGTWNLKDSKNFDEYMKALGVGFATRQVGGMTKPTTIIEVAGD
Cryodraco_antarcticus       --MVDVFVGTWNLKDSEKFDEYMKKLGVGFATRQVGNVTKPTTIISVEGD
Anguilla_japonica           MVIMEPFLGTWHLKTSENFDEYMKELGVGFATRKIGNTTKPTLIIAADGD
Fundulus_heteroclitus       --MVEAFVGTWNLKESENFDDYMKELGVGFATRKVGNLTKPTTIICVDGD
Tetraodon_nigroviridis      --MAEAFAGTWNLVKSEKFDEYMKELGVGLAMRKMGNLAKPTLSITIEGD
Rattus_norvegicus           --MCDAFVGTWKLVSSENFDDYMKEVGVGFATRKVAGMAKPNLIISVEGD
Danio_rerio                 --MVDKFVGTWKMTTSDNFDEYMKAIGVGFATRQVGNRTKPNLVVCVDEQ
Echinococcus_granulosus     ---MEAFLGTWKMEKSEGFDKIMERLGVDFVTRKMGNLVKPNLIVTDLGG
Taenia_solium               ---MEPFIGTWRMEKSEGFDKIMERLGVDFVTRKMGNLMKPSLIVSDLGD
Xenopus_laevis              --MVDQFVGSWKLTDSQGFDEYMQSLGVGFATRKAGAMAKPNVIISVNGD
                            xx45535454555545552445545525552555255555555545555

Salmo_salar                 -TVTLKTQSTFKNTEISFKLGEEFDETTADDRKVKSLITIDGGKMVHVQK
Cryodraco_antarcticus       -KVTLKTQSAIKNTELSFKLDEEFDETTADDRKVKSFVTLDGGKLVHTQK
Anguilla_japonica           -KFQVKTQSLLKSTEINFKLGEEFDETTADDRKVKSVVKLEDGKLVHLQK
Fundulus_heteroclitus       -KVTVKTQSTIKNTELSFKLGEEFDETTADDRKVKSLVTIEDGKLVHVQK
Tetraodon_nigroviridis      -KVTLKNSSTFKTTEVSFKLGEEFDESTADGRNVKSVVTVEDGKLVHVQK
Rattus_norvegicus           -LVVIRSESTFKNTEISFKLGVEFDEITPDDRKVKSIITLDGGVLVHVQK
Danio_rerio                 GLICMKSQSTFKTTEIKFKLNEPFEETTADDRKTTTVMTIENGKLVQKQT
Echinococcus_granulosus     GKYKMRSESTFKTTECSFKLGEKFKEVTPDSREVASLITVENGVMKHEQD
Taenia_solium               GKYSMRSESKFKTTEFTFKLGEKFKETTPDSREVTSLITVENGVMKQEQV
Xenopus_laevis              -EILLKTESSLKTTEVTFKLGQEFDEQTADNRKTKTIITCDSGVLNQVQK
                            x552555555355555551555534555425515553154542545544 5

Salmo_salar                 WDGKETTLVREVSGNALERT
Cryodraco_antarcticus       WDGKETSLVREVNGNSLTLT
Anguilla_japonica           WDSKETSLVRAVDGNKLTLT
Fundulus_heteroclitus       WDGKETTLVREVDGNKLTLT
Tetraodon_nigroviridis      WDGKETSLVREVEGNNLTLT
Rattus_norvegicus           WDGKSTTIKRRXDGDKLVVE
Danio_rerio                 WDGKESTIEREVSDGKLIAK
Echinococcus_granulosus     DKTKVTYIERVVEGNELKAT
Taenia_solium               GKGKTTYIDRVIEGNELKTT
Xenopus_laevis              WDGKETTIQREIKNGHLVVT
                            24555555552545555153
```

Figure S1 - Multisequence alignment of the fatty acid binding protein family. Predicted RCRRs (red) and experimentally examined regions (green) are marked. The rare codon score S is displayed for each column.

Table S1 - Organisms and GenBank identifiers of the fatty acid binding protein family.

Salmo_salar	GenBank: AAR91708
Cryodraco antarcticus	GenBank: AAC60357
Anguilla japonica	GenBank: BAA92355
Fundulus heteroclitus	GenBank: AAK61550
Tetraodon nigroviridis	GenBank: CAG10013
Rattus norvegicus	GenBank: NP_445817
Danio rerio	GenBank: NP_001004682
Echinococcus granulosus	GenBank: Q02970
Taenia solium	GenBank: ABB76135
Xenopus laevis	GenBank: AAH78499

```
Mannheimia_haemolytica        -----MNYTKFDVKNWVRREHFEFYRHRLPCGFSLTSKIDITTLKKSLDD
Aeromonas_salmonicida         -----MNFTRIDLNTWNRREHFAFYRQQIKCGFSLTTKLDITALRTALAE
Photorhabdus_luminescens      -----MNYSKVDIDLWDRKEHFLHYRNVVQCGFSLTAKIDITHLLSSLVE
Yersinia_pestis_biovar        MEKKITGYTTVDISQWHRKEHFEAFQSVAQCTYNQTVQLDITAFLKTVKK
Salmonella_typhimurium        ----------------------------------NQTVQLDITAFLKTVKK
Neisseria_meningitidis        -----MVFEKIDKNSWNRKEYFDHYFASVPCTYSMTVKVDITQIK----E
Streptococcus_suis            -----MNFNKIDLDNWKRKEIFNHYLN-QQTTFSITTEIDISVLYRNIKQ
Enterococcus_faecium          -----MTFNIINLETWDRKEYFNHYFN-QQTTYSVTKELDITLLKSMIKD
                              xxxxx3145555544555515255454554345555555553455423 45

Mannheimia_haemolytica        SAYKFYPVMIYLIAQAVNQFDELRMAIK-DDELIVWDSVDPQFTVFHQET
Aeromonas_salmonicida         TGYKFYPLMIYLISRAVNQFPEFRMALK-DNELIYWDQSDPVFTVFHKET
Photorhabdus_luminescens      KQYKFYPTMIYLISTVVNSYSEFRMAIK-DEELIVWDGVNPAYTIFHKET
Yersinia_pestis_biovar        NKHKFYPAFIHILARLMNAHPEFRMAMK-DGELVIWDSVHPCYTVFHEQT
Salmonella_typhimurium        NKHKFYPAFIHILARLMNAHPEFRMAMK-DGELVIWDSVHPCYTVFHEQT
Neisseria_meningitidis        KGMKLYPAMLYYIAMIVNRHSEFRTAINQDGELGIYDEMIPSYTIFHNDT
Streptococcus_suis            KRYKFYPAFVFLVTRVINSNTAFRTGYNSEGELGYWDKLDPLYTIFDSVS
Enterococcus_faecium          KGYELYPALIHAIVSVINRNKVFRTGINSEGNLGYWDKLEPLYTVFNKET
                              22551555554555551254543551 45x25453135431452514535 3

Mannheimia_haemolytica        ETFSALSCPYSSDIDQFMVNYLSVMERYKSDTKLFPQGVTPENHLNISAL
Aeromonas_salmonicida         ETFSALSCRYFPDLSEFMAGYNAVTAEYQHDTRLFPQGNLPENHLNISSL
Photorhabdus_luminescens      ETFSAIWTEFNSDLAEFMKNYSADYETYKDDLCFFSKPELPENHFHISSV
Yersinia_pestis_biovar        ETFSSLWSEYHDDFRQFLHIYSQDIACYGENLAYFPKG-FIENMFFVSAN
Salmonella_typhimurium        ETFSSLWSEYHDDFRQFLHIYSQDVACYGENLAYFPKG-FIENMFFVSAN
Neisseria_meningitidis        ETFSSLWTECKSDFKSFLADYESDTQRYGNNHRMEGKPNAPENIFNVSMI
Streptococcus_suis            KTFSGIWTPARNDFKEFYDLYLSDVEKYNGSGKLFPKTPIPENAFSISII
Enterococcus_faecium          EKFSNIWTESNASFNSFYNSYKNDLFKYKDKNEMFPKKPIPENTVPISMI
                              5435555451235453514355354553554525525141553452415 4
```

```
Mannheimia_haemolytica        PWVNFDSFNLNVANFTDYFAPIITMAKYQQEGDRLLLPLSVQVHHAVCDG
Aeromonas_salmonicida         PWVSFDGFNLNITGNDDYFAPVFTMAKFQQEGDRVLLPVSVQVHHAVCDG
Photorhabdus_luminescens      PWVSFDGFNLNMASVMDYFPPIFTMGKFYQNGNQTQLPLAIQVHHATCDG
Yersinia_pestis_biovar        PWVSFTSFDLNVANMDNFFAPVFTMGKYYTQGDKVLMPLAIQVHHAVCDG
Salmonella_typhimurium        PWVSFTSFDLNVANMDNFFAPVFTMGKYYTQGDKVLMPLAIQVHHAVCDG
Neisseria_meningitidis        PWSTFDGFNLNLQKGYDYLIPIFTMGKYYKEDNKIILPLAIQVHHAVCDG
Streptococcus_suis            PWTSFTGFNLNINNNSNYLLPIITAGKFINKGNSIYLPLSLQVHHSVCDG
Enterococcus_faecium          PWIDFSSFNLNIGNNSRFLLPIITIGKFYSKDDKIYLPFSLQVHHAVCDG
                              55555535354555545554153535554545543545555445555355


Mannheimia_haemolytica        FHVARFISRLQELCNSKLK---
Aeromonas_salmonicida         FHAARFINTLQLMCDNILK---
Photorhabdus_luminescens      FHVGRVINNLQELCNDFI----
Yersinia_pestis_biovar        FHVGRMLNELQQYCDEWQGGA-
Salmonella_typhimurium        LHVGRMLNELQQYCDEWQ----
Neisseria_meningitidis        FHICRFVNELQELINS------
Streptococcus_suis            YHAGLFMNSIQELADRPNDWLF
Enterococcus_faecium          YHVSLFMNEFQNIIDNVNEWI-
                             1545554445453242xxxxxx
```

Figure S2 - Multisequence alignment of the chloramphenicol acetyltransferase protein family. Predicted RCRRs (red) and experimentally examined regions (green) are marked. The rare codon score S is displayed for each column.

Table S2 - Organisms and GenBank identifiers of the chloramphenicol acetyltransferase protein family.

Mannheimia haemolytica	GenBank: NP_073222
Aeromonas salmonicida	GenBank: CAD57199
Photorhabdus luminescens	GenBank: NP_929686
Yersinia pestis biovar	GenBank: ZP_01174255
Salmonella typhimurium	GenBank: ABA56511
Neisseria meningitidis	GenBank: AAC14400
Streptococcus suis	GenBank: BAC11901
Enterococcus faecium	GenBank: NP_863168

Table S3 - Number of columns per family, sorted by groups S1 to S5, depending on their score S. Groups: S1 with ($0 \leq S < 0.2$), S2 with ($0.2 \leq S < 0.4$), S3 with ($0.4 \leq S < 0.6$), S4 with ($0.6 \leq S < 0.8$) and S5 with ($0.8 \leq S \leq 1$).

LED ID	1	2	3	4	5
abH01.02	29	47	60	93	336
abH08.14	4	8	20	35	143
abH09.02	0	13	16	51	173
abH12.01	6	20	26	51	157
abH14.02	8	17	38	71	273
abH15.02	18	22	51	74	149
abH17.01	11	14	22	44	140
abH19.01	17	29	52	55	146
abH23.01	4	8	24	48	219
abH24.01	15	20	33	41	502
abH26.01	6	10	29	53	227
abH28	6	8	19	68	602
abH30	1	6	20	43	522
abH31.02	0	4	5	17	208
abH33.01	4	14	16	47	272
abH34.02	26	28	59	105	268
Chloramphenicol acetyltransferase protein family	9	17	16	51	117
Fatty acid binding protein family	4	9	5	17	82
Total	168	294	511	964	4536

8.2. The isoelectric region of proteins: a systematic analysis

Michael Widmann,[1] Peter Trodler, Jürgen Pleiss[1§]

[1] Institute of Technical Biochemistry, Allmandring 31, 70569 Stuttgart

Publikation erschienen bei *PLoS ONE* 2010, 5(5)

8.2.1. Abstract

Background:

Binding of proteins in ion exchange chromatography is dominated by electrostatic interactions and can be tuned by adjusting pH and ionic strength of the solvent. Therefore, the isoelectric region (IER), the pH region of almost zero charge near the pI, has been used to predict the binding properties of proteins.

Principal findings:

Usually the IER is small and binding and elution is carried out at pH values near to the pI . However, some proteins with an extended IER have been shown to bind and elute far away from its pI. To analyze factors that mediate the size of the IER and to identify proteins with an extended IER, two protein families consisting of more than 7000 proteins were systematically investigated. Most proteins were found to have a small IER and thus are expected to bind or elute near to their pI, while only a small fraction of less than 2% had a large IER.

Conclusions:

Only four factors, the number of histidines, the pI, the number of titratable amino acids and the ratio of acidic to basic residues, are sufficient to reliably classify proteins by their IER based on their sequence only, and thus to predict their binding and elution behaviour in ion exchange chromatography.

Keywords

Electrostatic potential; ion exchange chromatography; titration curve; isoelectric region, database analysis

8.2.2. Introduction

Ion exchange chromatography (IEC) is a widely applied method in protein purification. It is well established, efficient, and applicable to large scale purification (Palekar, Vasudevan et al. 2000; Ahamed, Ottens et al. 2006). Protein binding in IEC is primarily determined by electrostatic interactions between the charge of the protein and the charged stationary phase (Sheehan and FitzGerald 1996; Hallgren, Kalman et al. 2000; Sheehan and O'Sullivan 2001). As a consequence, optimal pH values for binding to or elution from an ion exchange column can be predicted by the isoelectric point (pI) for many proteins (Ahamed, Nfor et al. 2007) with loading pHs about 0.5-1 pH units above or below the pI of the respective protein (Healthcare 2004; Ahamed, Chilamkurthi et al. 2008). However, it has been shown that for some proteins the pI is not predictive, but binding to or elution from the column only occurs for pH values far from the pI of the protein (Ahamed, Nfor et al. 2007; Trodler, Nieveler et al. 2008). A detailed investigation of pH values at which bound proteins eluted from an anion exchange chromatography column were performed using a pH gradient as the method of elution. It demonstrated that for proteins with pI values between 6 and 8, the elution occurred at pH values considerably higher than their pI. Proteins with a pI lower than 6 or higher than 8 however eluted at pH values close to their pI (Ahamed, Nfor et al. 2007). The unique behaviour of proteins with a pI between 6 and 8 was explained by the observation that their titration curves had a broad region of almost zero charge near their pI which extended over several pH units (Ahamed, Nfor et al. 2007). In this work, we term this region the isoelectric region (IER). A large IER has also been shown to influence the binding of proteins to ion exchange columns which has been demonstrated for the lipase B from *Candida antarctica* (Trodler, Nieveler et al. 2008). The purification of this protein by ion exchange chromatography had not been achieved before. Only by taking the large IER into account and substantially lowering the binding pH to 3, which is 3 pH units lower than the pI of the protein and beyond the proteins IER, a successful binding to a cation exchange column was achieved. Therefore, a large IER is expected to lead to differences between the pI of a protein and its pH of binding to or elution from a column. Two factors which can be easily extracted from the protein sequence have been suggested to determine a large IER: a pI of the protein between 6 and 8 (Ahamed, Nfor et al. 2007), and a low number of histidines (Trodler, Nieveler et al. 2008), since histidine is the only titratable residue in the pH region between 5 and 9.

In this work, we investigated these factors by a systematic analysis of two protein families, the α/β hydrolase family with more than 4600 proteins and the medium-chain dehydrogenase/reductase protein family with more than 2600 proteins, based on the Lipase Engineering Database (Fischer and Pleiss 2003) and the Medium-Chain Dehydrogenase/Reductase Engineering Database (Knoll and Pleiss 2008), respectively. Both protein families had previously been integrated in our data warehouse system for protein families DWARF (Fischer, Thai et al. 2006). The members of each protein family share a similar structure but have highly diverse sequences. In addition, the results were compared to a set of 5000 randomly generated protein sequences. The frequency of proteins with a large IER and the influence of the previously suggested factors like the number of histidines and the pI on the IER were investigated in order to establish a set of factors with a correlation to the IER. The ratio of acidic and basic amino acids R was included as a factor for this analysis since it had been previously shown to correlate with the pI (Patrickios and Yamasaki 1995). These factors could be used to change the IER by protein engineering in order to facilitate the purification process by ion exchange chromatography methods. To allow for a direct access to data on the isoelectric point and the size of the IER, these values were pre-calculated and integrated in our database model.

8.2.3. Results

Isoelectric region

The size of the region of very low total charge (larger than -3 and smaller than 3) near the pI of a protein differs significantly between proteins and was termed the isoelectric region (IER) in this work. Proteins with a small IER are expected to bind to or elute from an ion exchange column at a pH value close to their pI because their total charge sensitively depends on the pH at values close to their pI. Proteins with a large IER, however, are expected to bind to or elute from an ion exchange column at pH values that are noticeably higher or lower than their pI due to the elongated area of almost zero charge in proximity to the pI. To determine the number of proteins with a large IER, the IER of 4652 sequences from the α/β hydrolases family and 2683 sequences from the medium-chain dehydrogenase/reductase protein family were evaluated and systematically analyzed and compared to 5000 random sequences. The calculated IER ranged from 0.1 to 5.2. The proteins were divided into 2 groups depending on their IER: proteins with a small IER ($0.1 \leq IER < 3$) and proteins with a large IER ($3 \leq IER$). For both protein families and the random set the majority of proteins (98%) belonged to the first group with a small IER. The distribution of proteins in regard to the IER was found to be identical for all three protein sets and only a small minority of proteins in each set (2%) belonged to the group with a large IER (Figure 1). Only a few protein families constituted the group with a large IER. For the α/β hydrolase family, 40% of proteins with a large IER were members of the ‘cutinase’, ‘antigen 85’, or ‘carboxylesterase’ families. For the medium-chain dehydrogenase/reductase protein family, 70% of all proteins with a large IER belonged to the ‘YADH’ or the ‘QOR like’ families. However, the majority of proteins in these protein families also had a small IER.

Factors influencing the IER

Several factors were investigated for their correlation with the size of the IER. The goal was to identify a single factor or a combination of factors which showed a strong correlation with the IER. Factors that were considered for this analysis were the number of histidines, the pI of the protein, and the related ratio R between acidic and basic amino acids. These factors were chosen since they had already been shown or suggested to influence a proteins IER.

The number of histidines and the size of the IER were determined for every protein. For proteins with an identical number of histidines, the mean IER and its standard deviation were calculated. If only 2 or less proteins had the same number of histidines they were excluded

from the analysis. The number of histidines showed only a weak correlation to the size of the IER for the α/β hydrolase and medium-chain dehydrogenase/reductase families, and no correlation for the random set (Supporting Information 1, Figure S1). Furthermore, proteins with the same number of histidines showed considerable differences in the size of the IER indicated by large standard deviations (e.g. ±1.5 for proteins with 2 histidines). The number of histidines, independent of other factors, was therefore shown to be an inadequate factor for the prediction of the IER.

The other previously suggested factor to be indicative for proteins with a large IER was a pI value between 6 and 8. Therefore the pI was determined for each protein and the proteins were divided into two groups based on their pI. One group consisted of proteins with a pI between 6 and 8, the other group of proteins with a pI lower than 6 or higher than 8. Proteins with a pI between 6 and 8 were shown to have a higher percentage of proteins with a large IER (3-5%) than proteins with a pI below 6 or above 8 (1%). This distribution was observed for both protein families and the random set.

The analysis of the correlation between the number of histidines and the size of the IER was repeated for the two protein groups that were assigned based on pI. Now, a strong dependence between the number of histidines and the size of the IER was observed for proteins with pI values between 6 and 8 for both protein families and the random set (Supporting Information 2, Figure S2). Proteins with the same number of histidines had a similar IER as indicated by small standard deviations and showed a steady decrease of their IER with an increasing number of histidines. For proteins with a pI lower than 6 or higher than 8 only a very weak correlation to the number of histidines could be observed and proteins in this group generally had a small IER. However, a few proteins in this group still showed a large IER indicated by the large standard deviations in this set.

To find a factor that identifies proteins with a small IER, independently of the number of histidines, the ratio R of acidic and basic amino acids was combined with the number of titratable residues, because we observed that all proteins with a large IER showed a balanced ratio R of acidic and basic amino acids and a low number of titratable residues, in contrast to the majority of proteins with a small IER. These two properties were combined into a new balance factor B by multiplying the number of titratable residues by |ln R| (Material and Methods). Analogous to the classification of proteins by their pI, the factor B was used to separate the proteins into two groups. Both groups were then evaluated for a correlation between the number of histidines and the size of the IER.

According to this evaluation, a threshold of 6 was selected for the factor B which yielded the best separation for both groups in regard to the correlation between the IER and the number of histidines. For proteins with $B \leq 6$, a strong dependence of the IER on the number of histidines was observed (Figure 2). For the α/β hydrolase family, the average IER for proteins with one histidine was 4.3. It decreased to 2.6 for proteins with 6 histidines, and to 1.6 for proteins with 18 histidines. This was similar to the average IER from the medium-chain dehydrogenase/reductase protein family which also showed an average IER of 2.6 for proteins with 6 histidines. The set of random sequences showed an average IER of 3.4 for proteins with two histidines and an IER of 2.5 for proteins with 6 histidines. In contrast, all proteins with $B > 6$ had a small IER of less than 3 (Tab. 1), showed no correlation of the IER with the number of histidines and displayed a median IER of 1.6 or less for proteins with the same number of histidines (Figure 2).

Database integration

The values for the pI, IER and the charge of the protein for pH values 0-14 were calculated and integrated in the database model of the Lipase Engineering Database (LED). The pI and the titration curve are directly accessible via the web interface and are displayed in tabular as well as in graphical form for the selected sequence (Figure 3). In addition, the size of the isoelectric region (IER) is calculated and displayed. The LED is accessible by a web interface at http://www.led.uni-stuttgart.de.

8.2.4. Discussion

The isoelectric region (IER) of a protein is known to have a considerable influence on the binding to or elution from an ion exchange column (Ahamed, Nfor et al. 2007; Trodler, Nieveler et al. 2008). This influence is based on the size of the region of almost zero charge near the pI of the protein. For proteins with a small IER, the optimal pH for binding or elution can usually be predicted by their pI. While this prediction often matches experimental results reasonably well, other factors besides the net charge can influence the binding behaviour of proteins to ion exchange columns. This includes the surface charge distribution (Hallgren, Kalman et al. 2000), protein hydrophobicity (Melander, el Rassi et al. 1989; Malmquist, Nilsson et al. 2006), van der Waals interactions (Roth and Lenhoff 1995), and choice of the adsorbent materials (Sheehan and FitzGerald 1996; Noh, Yohe et al. 2008). It has also been shown that not only the amino acid composition of a protein but also its subsequent modification can influence the elution behaviour of proteins in ion exchange chromatography, e.g. by glycosylation which might lead to the shielding of surface charges (Gotte, Libonati et al. 2003). For proteins with a large IER, however, the net charge in combination with the IER has been shown to be the major factor that influences binding to (Trodler, Nieveler et al. 2008) or elution from (Ahamed, Nfor et al. 2007) an ion exchange column. While the pI is a widely used parameter for the estimation of the electrostatic interactions of proteins, the prediction of the IER is frequently neglected. This can be explained by the small number (2%) of proteins having a large IER. However, for these proteins the importance of the IER has been demonstrated and should be taken into consideration in addition to the pI. The size of the IER can be easily determined from a calculated or experimentally determined titration curve.

In order to understand the molecular basis of a small or large IER, factors that correlate with the size of the IER were identified and analysed. This included previously suggested factors like a pI between 6 and 8 (Ahamed, Nfor et al. 2007) or a low number of histidines (Trodler, Nieveler et al. 2008). We could show that neither of these factors was correlating with the size of the IER on its own. However, by combining two factors into the balance factor B, proteins which showed a correlation of the IER to the number of histidines were identified. Proteins with a value of B less than 6 had a large IER if they included only a small number of histidines, while their IER decreased with increasing number of histidines. For these proteins, the number of histidines is not only a good indicator of the size of the pI, but histidine would also be the major target of engineering a variant with a changed IER. The study also showed that for many proteins the size of the IER sensitively depends on the number and ratio of

charged amino acids. Even a small number of amino acid exchanges in protein mutants or isoforms may therefore have a large impact on the optimal pH of binding to an ion exchange column and other charged surfaces.

The integration of the isoelectric point and the size of the IER in our database model of the LED furthermore allows for a direct access to these values and a visualization of the titration curve for each protein in the database.

8.2.5. Material and Methods

Titration curve calculation

Protein sequences were taken from the Lipase Engineering Database (Fischer and Pleiss 2003) and the Medium-Chain Dehydrogenase/Reductase Engineering Database (Knoll and Pleiss 2008). Sequences with 100% sequence identity and fragments with a length of less than 160 amino acids were excluded, resulting in a total of 4652 sequences from the α/β hydrolase family and 2683 sequences from the dehydrogenase/reductase protein family A set of 5000 random sequences was generated using frequencies for the titratable amino acids from (Mitra and Rani 1993) (Supporting Information 3, Tab. S1). The distribution of titratable amino acids was similar to the distribution found in the α/β hydrolase and dehydrogenase/reductase protein families (Supporting Information 3, Tab. S2, Tab. S3). The random set had a defined protein size range between 250 – 450 amino acids, similar to the size distribution of the dehydrogenase/reductase protein family. Protein charges were calculated using the module "pICalculator" from the Bioperl toolkit (Stajich, Block et al. 2002). 6 titratable amino acids were included: aspartate (Asp), glutamate (Glu), histidine (His), tyrosine (Tyr), lysine (Lys), and arginine (Arg); pK_a values were assigned as described previously (Rice, Longden et al. 2000): 3.9, 4.1, 6.5, 10.1, 10.8, and 12.5, respectively. The N- and C- termini had a pK_a of 8.6 and 3.6 respectively. Cysteine (Cys) was treated as a nontitratable residue because sequence-based methods are unable to distinguish between free cysteines and cysteines that are part of disulfide bridges. For 112 α/β hydrolases with experimentally determined structures, at least 65% of all cysteines were found to be part of disulfide bridges (data not shown). This number is supposed to be even higher because not all disulfide bridges are properly annotated in the structure entries. Previously it was found that 91% of all cysteines were part of disulfide bridges in over 50 analyzed proteins (Patrickios and Yamasaki 1995). In order to validate the accuracy of predictions calculated with the Emboss pKa set, a comparison of this set, a more recent pKa set (Grimsley, Scholtz et al. 2009), and a structure based method (PDB2PQR/PROPKA (Dolinsky, Czodrowski et al. 2007)) was performed. 25 proteins with resolved crystal structures were randomly chosen from the data set, and the amino acid sequences used for all calculations were extracted from the crystal structure file. For pH values between 1 to 14, the total charge of the proteins was calculated as the sum of the partial charges of each titratable group. The comparison demonstrated that for the sequence based methods the deviation between the predicted IER and pI values were less than 0.3 and 0.4, respectively (Supporting Information 4, Table S4). The deviation between the Emboss pKa

set and the structure based approach using PDB2PQR/PROPKA (Dolinsky, Czodrowski et al. 2007) was less than 0.6 for the IER and 0.8 for the pI (Supporting Information 4, Table S5, Figure S3).

Ratio between acidic and basic amino acids

Previously, it was shown that the pI of a protein correlates with the ratio R of acidic and basic amino acids:

$$R \equiv \frac{Asp + Glu}{Arg + Lys}$$

Proteins with a balanced ratio R ≈ 1 showed a high sensitivity of pI to R, while for proteins with an unbalanced ratio the pI was insensitive to R. Previously the ratio R was compared to experimentally determined pI values for 58 proteins (Patrickios and Yamasaki 1995).

Since a factor R of 1 implies a balance of acidic and basic amino acids, the absolute value of the logarithm of R is a good measurement for the imbalance between acidic and basic amino acids. The introduced balance factor B takes the total number of titratable residues of each protein into account in addition to the distribution of acidic and basic amino acids as represented by |ln R|. The total number of titratable residues was designated as T and is multiplied with |ln R|, resulting in the factor B.

$$B = \left|\ln R\right| * T$$

8.2.6. Acknowledgements

We thank Florian Wagner for programming of the dynamic user interface.

8.2.7. References

Ahamed, T., S. Chilamkurthi, et al. (2008). "Selection of pH-related parameters in ion-exchange chromatography using pH-gradient operations." Journal of Chromatography A **1194**(1): 22-29.

Ahamed, T., B. K. Nfor, et al. (2007). "pH-gradient ion-exchange chromatography: an analytical tool for design and optimization of protein separations." J Chromatogr A **1164**(1-2): 181-8.

Ahamed, T., M. Ottens, et al. (2006). "A generalized approach to thermodynamic properties of biomolecules for use in bioseparation process design." Fluid Phase Equilibria **241**(1-2): 268-282.

Fischer, M. and J. Pleiss (2003). "The Lipase Engineering Database: a navigation and analysis tool for protein families." Nucleic Acids Res **31**(1): 319-21.

Fischer, M., Q. K. Thai, et al. (2006). "DWARF--a data warehouse system for analyzing protein families." BMC Bioinformatics **7**: 495.

Hallgren, E., F. Kalman, et al. (2000). "Protein retention in ion-exchange chromatography: effect of net charge and charge distribution." J Chromatogr A **877**(1-2): 13-24.

Healthcare, G. (2004). Ion Exchange Chromatography & Chromatofocusing:Principles and Methods, GE Healthcare.

Knoll, M. and J. Pleiss (2008). "The Medium-Chain Dehydrogenase/reductase Engineering Database: a systematic analysis of a diverse protein family to understand sequence-structure-function relationship." Protein Sci **17**(10): 1689-97.

Malmquist, G., U. H. Nilsson, et al. (2006). "Electrostatic calculations and quantitative protein retention models for ion exchange chromatography." J Chromatogr A **1115**(1-2): 164-86.

Melander, W. R., Z. el Rassi, et al. (1989). "Interplay of hydrophobic and electrostatic interactions in biopolymer chromatography. Effect of salts on the retention of proteins." J Chromatogr **469**: 3-27.

Mitra, C. K. and M. Rani (1993). "Protein Sequences as Random Fractals." Journal of Biosciences **18**(2): 213-220.

Noh, H., S. T. Yohe, et al. (2008). "Volumetric interpretation of protein adsorption: Ion-exchange adsorbent capacity, protein pI, and interaction energetics." Biomaterials **29**(13): 2033-48.

Palekar, A. A., P. T. Vasudevan, et al. (2000). "Purification of lipase: A review." Biocatalysis and Biotransformation **18**(3): 177-200.

Patrickios, C. S. and E. N. Yamasaki (1995). "Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory." Anal Biochem **231**(1): 82-91.

Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Roth, C. M. and A. M. Lenhoff (1995). "Electrostatic and Van-Der-Waals Contributions to Protein Adsorption - Comparison of Theory and Experiment." Langmuir **11**(9): 3500-3509.

Sheehan, D. and R. FitzGerald (1996). "Ion-exchange chromatography." Methods Mol Biol **59**: 145-50.

Sheehan, D. and S. O'Sullivan (2001). "Ion Exchange Chromatography." Encyclopedia of Life Sciences: -.

Stajich, J. E., D. Block, et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-8.

Trodler, P., J. Nieveler, et al. (2008). "Rational design of a new one-step purification strategy for Candida antarctica lipase B by ion-exchange chromatography." Journal of Chromatography A **1179**(2): 161-1

8.2.8. Figures

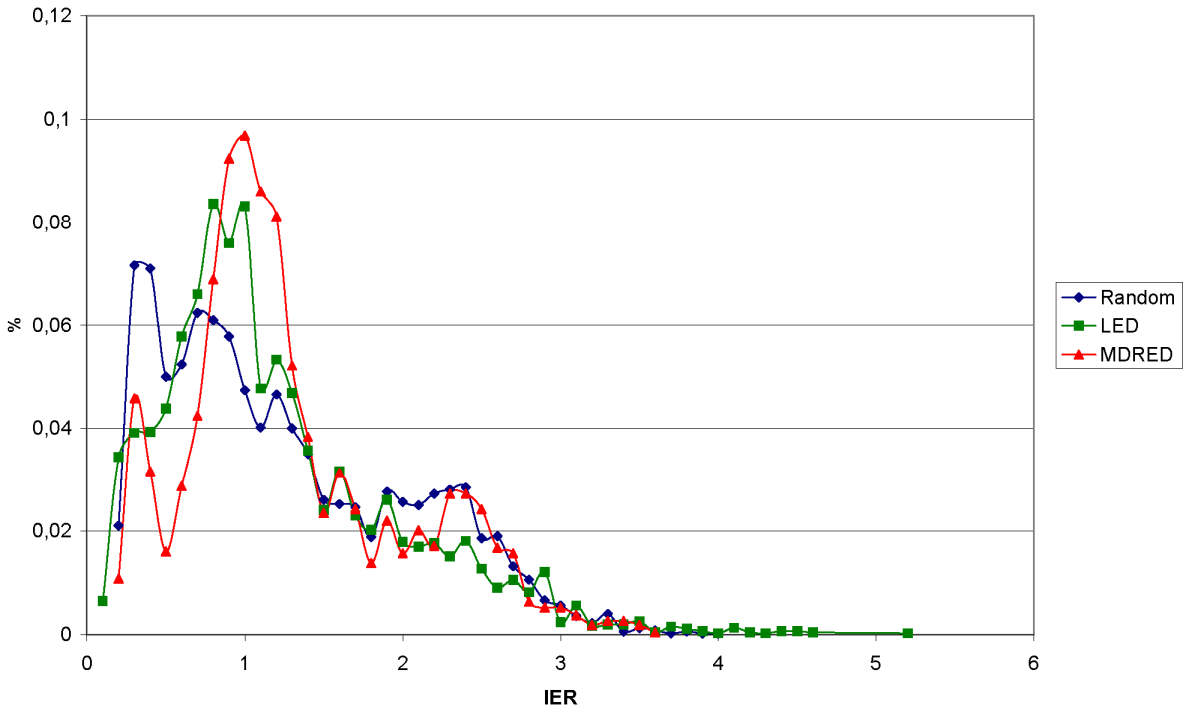


Figure 1 - Comparison of protein distribution according to IER size.

Protein numbers are displayed as percentages. Results from the α/β hydrolase database (LED) and the medium-chain dehydrogenase/reductase protein family (MDRED) are overlaid with results from the random set.

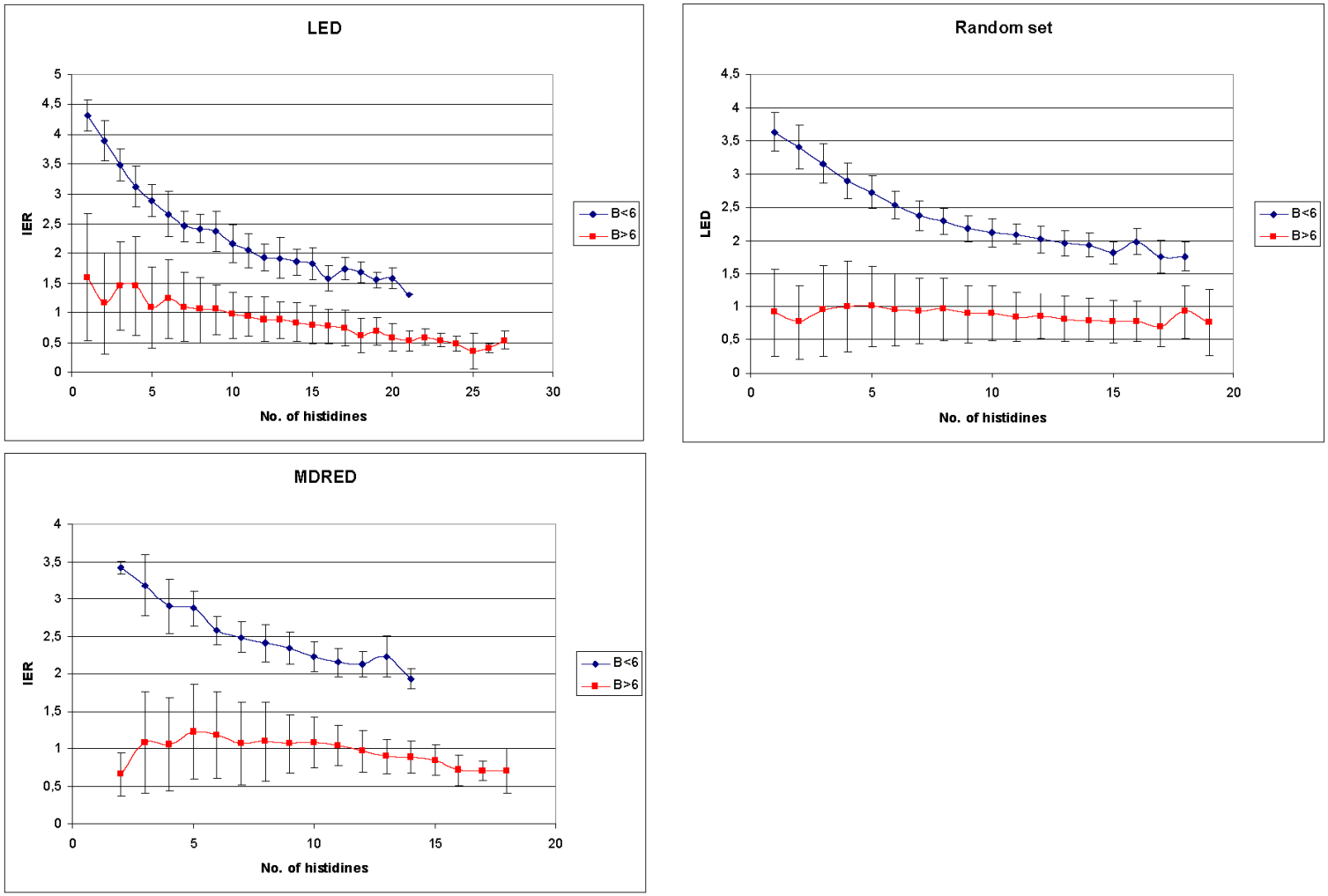


Figure 2 - Number of histidines and isoelectric region (IER) for protein family groups. For proteins with the same number of histidines, the median IERs are plotted against the number of histidines. Proteins with ($B \leq 6$) are depicted in blue, proteins with ($B > 6$) are depicted in red.

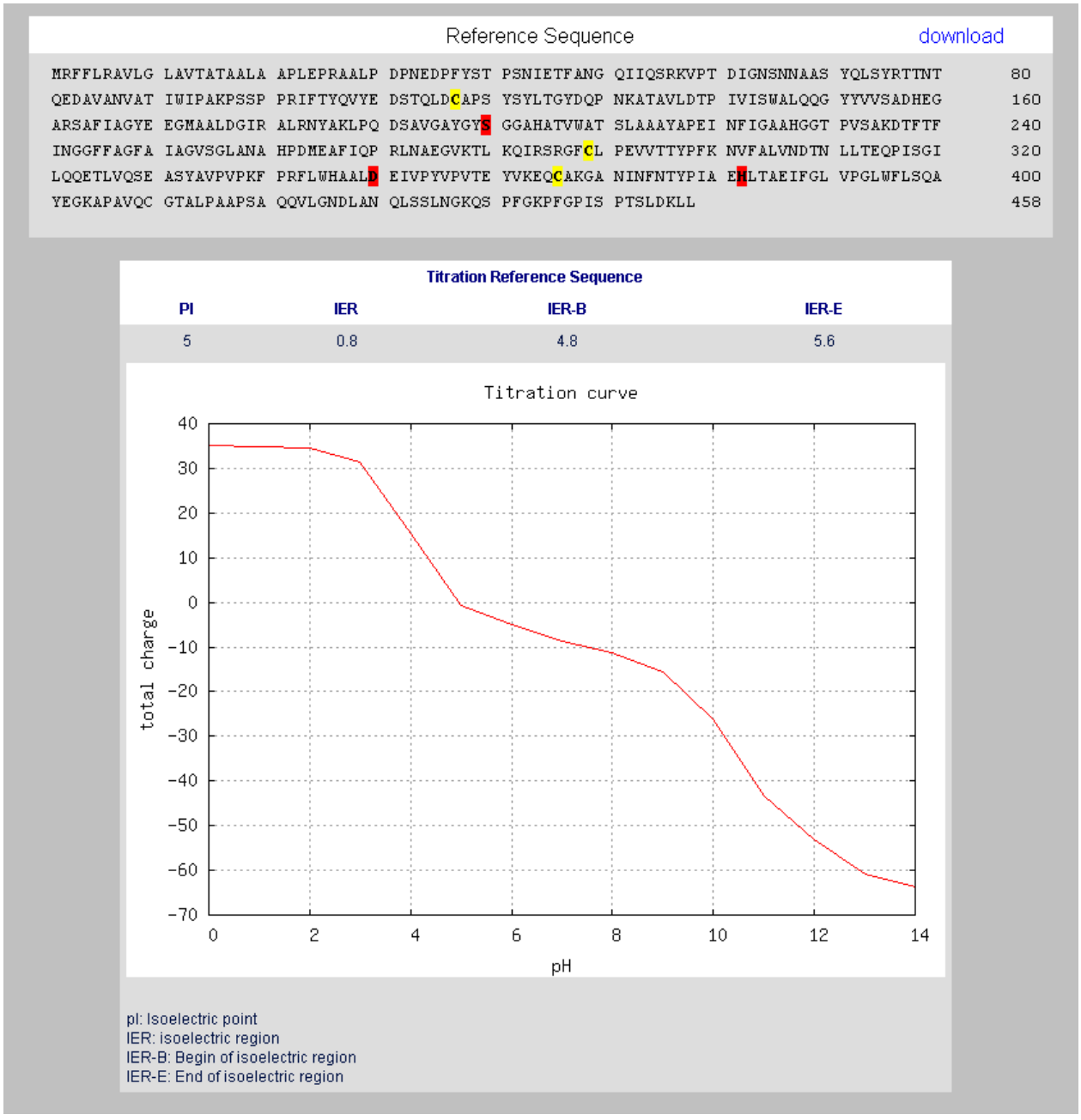


Figure 3 - Web interface of the LED with the electrostatic properties feature.

Protein total charge is displayed for each pH value from 0-14 in graphical form. Graphical representations of titration curves are generated by GNUPLOT. The isoelectric point and the size of the isoelectric region are given in tabular form.

8.2.9. Tables

Table 1 - Distribution of proteins according to IER size in dependency of B.

Protein family	(IER ≥ 3)	
	B <= 6	B > 6
α/β hydrolases	13%	0%
dehydrogenases/reductases	10%	0%
Random set	8%	0%

8.2.10. Supporting Information

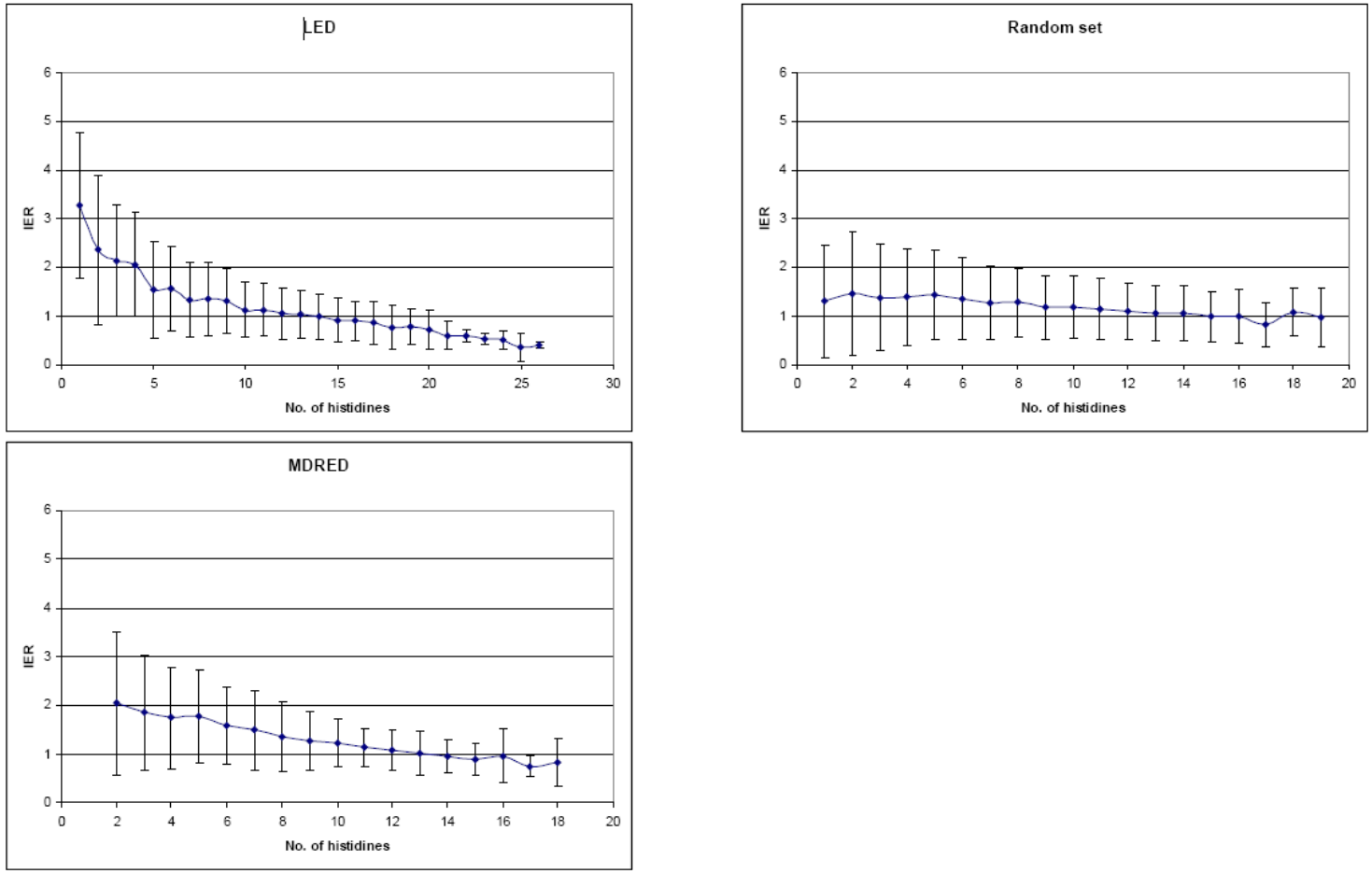


Figure S1 - Number of histidines and isoelectric region (IER) for each protein family.

For proteins with the same number of histidines, the median IERs are plotted against the number of histidines.

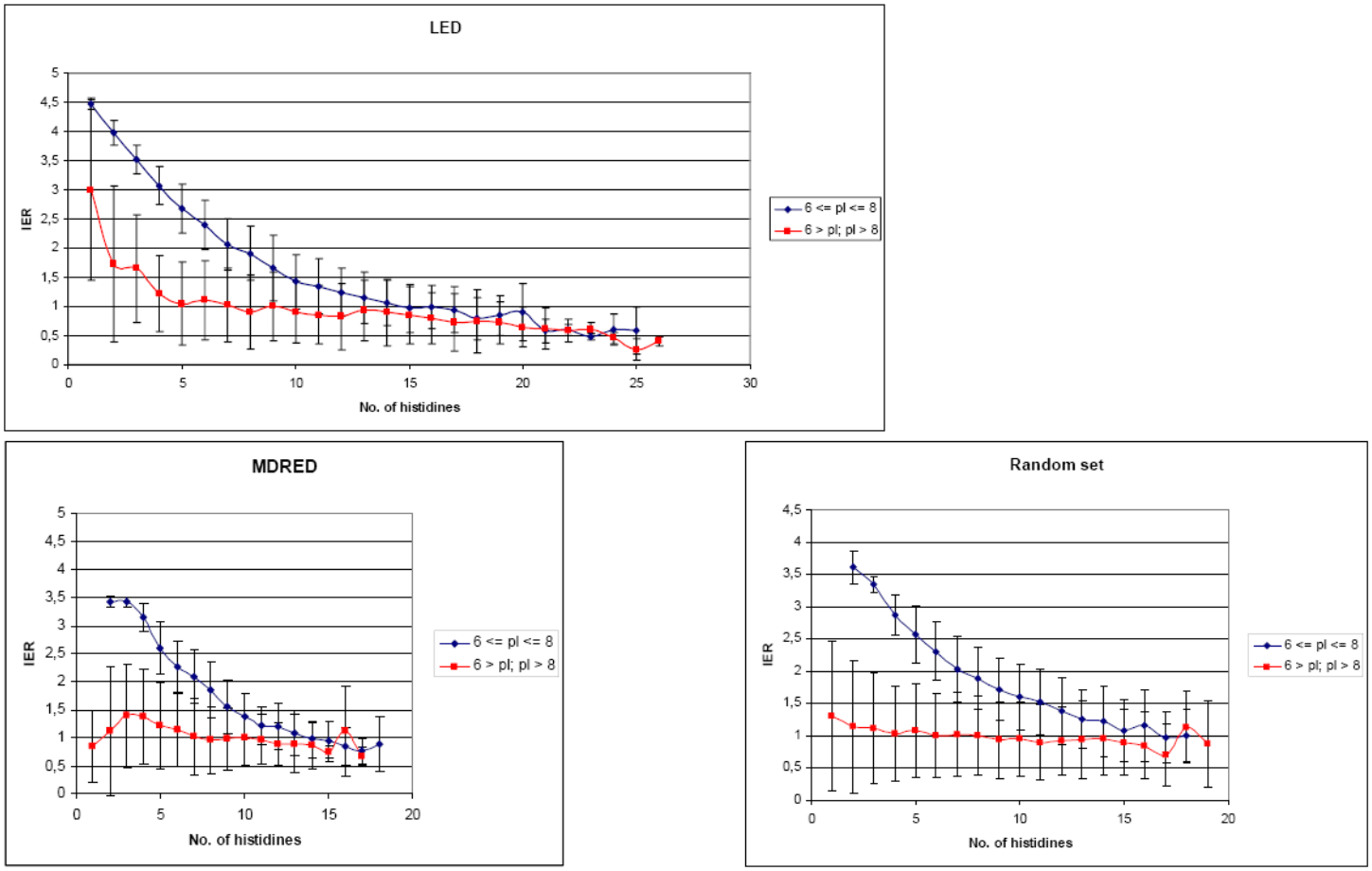


Figure S2 - Number of histidines and isoelectric region (IER) for each protein family depending on the pI. For proteins with the same number of histidines, the median IERs are plotted against the number of histidines. Proteins with (6 <= pI <= 8) are depicted in blue, proteins with (6 > pI; pI > 8) are depicted in red.

Table S1

Probabilities of titratable amino acids in percentages used for the creation of random sequences.

Aspartate (Asp)	5.2
Glutamate (Glu)	6.2
Histidine (His)	2.3
Tyrosine (Tyr)	3.2
Lysine (Lys)	5.8
Arginine (Arg)	5.2

Table S2

Distribution of titratable amino acids in percentages for all proteins of the α/β hydrolase family

Aspartate (Asp)	5.6
Glutamate (Glu)	5.2
Histidine (His)	2.7
Tyrosine (Tyr)	3.9
Lysine (Lys)	4.6
Arginine (Arg)	4.7

Table S3

Distribution of titratable amino acids in percentages for all proteins of the dehydrogenase/reductase family

Aspartate (Asp)	5.4
Glutamate (Glu)	6
Histidine (His)	2.6
Tyrosine (Tyr)	2.3
Lysine (Lys)	5.1
Arginine (Arg)	4.5

Table S4

Comparison of the calculated values for the IER and the pI of 25 proteins. Sequences were extracted from the crystal structure file given for each protein. Values were calculated with the Emboss pKa set used for all calculations in this work and more recent pKa values from (Grimsley, Scholtz et al. 2009).

PDB identifier	Emboss pKa set		pKa set by Grimsley et al 2009	
	IER	pI	IER	pI
1A7U	0.4	4.5	0.5	4.4
1AKN	0.9	5.7	1.1	5.7
1BN6	0.4	4.9	0.5	4.8
1CQW	0.4	4.9	0.5	4.8
1CRL	0.2	4.6	0.4	4.4
1CUB	4.5	7.5	4.6	7.2
1EVE	0.8	6.2	0.8	6.2
1EVQ	0.5	4.9	0.6	4.9
1EXW	1.9	6.8	1.6	6.8
1GKK	0.9	6.5	0.8	6.5
1I6W	1.1	9.7	1	9.7
1IUO	0.6	4.9	0.7	4.9
1J1I	1.2	6.2	1.3	6.3
1JKM	0.2	4.6	0.3	4.5
1JU3	0.2	4.5	0.3	4.2
1KU0	1.3	6.8	1.1	6.9
1L7R	0.2	4.4	0.2	4.2
1LGY	1.7	9.2	1.8	9.2
1MAA	1	6	1	6.1
1ODS	1.2	5.7	1.2	5.7
1OXM	4.5	8.8	4.7	8.4
1QJ4	0.8	5.1	1	5.1
1TCA	4.3	6	4.5	5.9
1XZA	4.5	8.8	4.7	8.4
2CUT	4.5	8.8	4.7	8.4

Table S5

Comparison of the calculated values for the IER and the pI of 25 proteins. Sequences were extracted from the crystal structure file given for each protein. Values were calculated with the Emboss pKa set and compared to the results of a structure based prediction performed with PDB2PQR/PROPKA (Dolinsky, Czodrowski et al. 2007) with the Parse force field.

PDB identifier	Emboss pKa set		PDB2PQR/PROPKA	
	IER	pI	IER	pI
1A7U	0.4	4.5	0.5	3.8
1AKN	0.9	5.7	1	5.7
1BN6	0.4	4.9	0.5	4.5
1CQW	0.4	4.9	0.5	4.5
1CRL	0.2	4.6	0.6	4.3
1CUB	4.5	7.5	4.5	6.7
1EVE	0.8	6.2	1.1	6.6
1EVQ	0.5	4.9	0.6	4.8
1EXW	1.9	6.8	1.9	7.1
1GKK	0.9	6.5	1.5	6.3
1I6W	1.1	9.7	1.1	9.4
1IUO	0.6	4.9	0.5	4.4
1J1I	1.2	6.2	1.6	5.9
1JKM	0.2	4.6	0.5	4.2
1JU3	0.2	4.5	0.3	3.9
1KU0	1.3	6.8	1.7	7.3
1L7R	0.2	4.4	0.2	4
1LGY	1.7	9.2	1.9	9.2
1MAA	1	6	1.3	6.1
1ODS	1.2	5.7	0.8	5.4
1OXM	4.5	8.8	4.5	8.6
1QJ4	0.8	5.1	0.6	4.6
1TCA	4.3	6	4	6.8
1XZA	4.5	8.8	4.5	8.6
2CUT	4.5	8.8	4.4	8.3

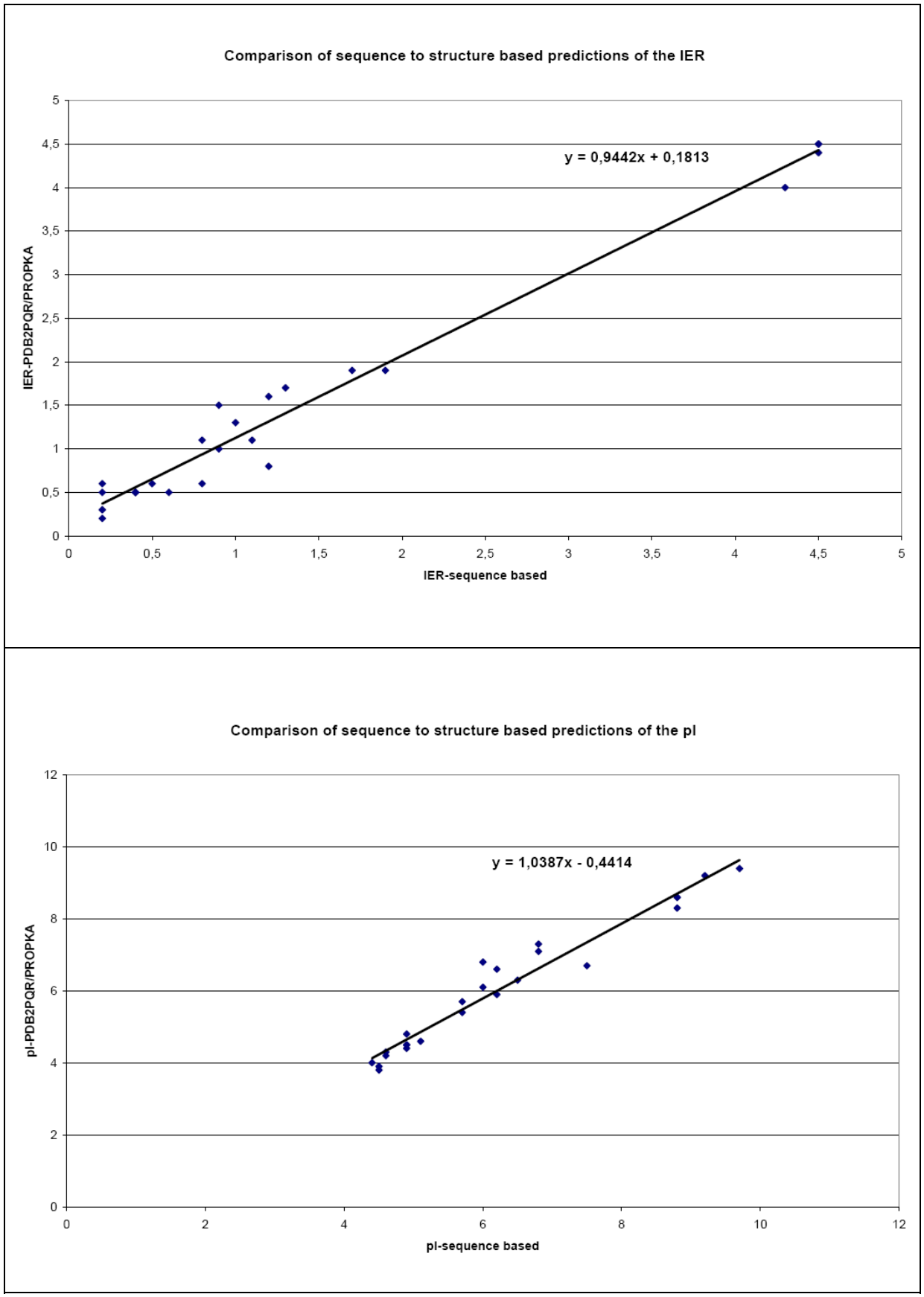


Figure S3 – Comparison of sequence to structure based predictions of the IER and the pI

Values for the IER and the pI were taken from Table S5.

8.3. Structural classification by the Lipase Engineering Database: a case study of *Candida antarctica* lipase A

Michael Widmann,[1] Benjamin Juhl, [1] and Jürgen Pleiss [1]

[1] Institute of Technical Biochemistry, University of Stuttgart,
Allmandring 31, 70569 Stuttgart, Germany

Publikation erschienen bei *BMC Genomics* 2010, 11:123

8.3.1. Abstract

Background

The Lipase Engineering Database (LED) integrates information on sequence, structure and function of lipases, esterases and related proteins with the α/β hydrolase fold. A new superfamily for *Candida antarctica* lipase A (CALA) was introduced including the recently published crystal structure of CALA. Since CALA has a highly divergent sequence in comparison to other α/β hydrolases, the Lipase Engineering Database was used to classify CALA in the frame of the already established classification system. This involved the comparison of CALA to similar structures as well as sequence-based comparisons against the content of the LED.

Results

The new release 3.0 (December 2009) of the Lipase Engineering Database contains 24783 sequence entries for 18585 proteins as well as 656 experimentally determined protein structures, including the structure of CALA. In comparison to the previous release (Fischer, Thai et al. 2006) with 4322 protein and 167 structure entries this update represents a significant increase in data volume. By comparing CALA to representative structures from all superfamilies, a structure from the deacetylase superfamily was found to be most similar to the structure of CALA. While the α/β hydrolase fold is conserved in both proteins, the major difference is found in the cap region. Sequence alignments between both proteins show a sequence similarity of only 15%. A multisequence alignment of both protein families was used to create hidden Markov models for the cap region of CALA and showed that the cap region of CALA is unique among all other proteins of the α/β hydrolase fold. By specifically comparing the substrate binding pocket of CALA to other binding pockets of α/β hydrolases, the binding pocket of *Candida rugosa* lipase was identified as being highly similar. This similarity also applied to the lid of *Candida rugosa* lipase in comparison to the potential lid of CALA.

Conclusion

The LED serves as a valuable tool for the systematic analysis of single proteins or protein families. The updated release 3.0 was used for the evaluation of α/β hydrolases. The HTML version of the database with new features is available at http://www.led.uni-stuttgart.de and

provides sequences, structures and a set of analysis tools including phylogenetic trees and HMM profiles

8.3.2. Background

Lipases (triacylglycerol hydrolases E.C. 3.1.1.3) are a versatile group of enzymes which catalyze the hydrolysis or synthesis of a broad range of water insoluble esters.

They belong to the class of α/β-hydrolases which also contains esterases, acetylcholinesterases, cutinases, carboxylesterases and epoxide hydrolases. Despite their high diversity in sequence and function, the α/β-hydrolases share a common architecture, the α/β-hydrolase fold (Ollis, Cheah et al. 1992) and conserved active site signatures, the GxSxG and GxDxG motifs (Pleiss, Fischer et al. 2000; Barth, Fischer et al. 2004). Two conserved features found in all α/β-hydrolases are the active site, consisting of the catalytic triad of S-D(E)-H, and the oxyanion hole. Depending on the amino acids involved in forming the oxyanion hole, the enzymes can be classified into three classes, the GGGX-, GX-, and the Y-class (Pleiss, Fischer et al. 2000). The Lipase Engineering Database (LED) (Fischer and Pleiss 2003) is a resource of fully and consistently annotated superfamilies and homologous families of α/β hydrolases including multisequence alignments of all families. The curation and annotation process for the LED is supported by DWARF (Fischer, Thai et al. 2006), an inhouse data warehouse system for protein families. The LED is accessible by a web interface at http://www.led.uni-stuttgart.de. It can be browsed on the level of families, organisms, or structures, and BLAST searches can be performed against all sequence entries.

Prominent members of the α/β hydrolases are the two lipases from *Candida antarctica*. Lipase B is a versatile and well characterized biocatalyst in many organic syntheses and biotransformations (Orrenius, Norin et al. 1995; Orrenius, Ohrner et al. 1995; Gotor-Fernandez, Busto et al. 2006) and shows a low sequence similarity to other α/β hydrolases. The second lipase from *Candida antarctica*, lipase A (CALA), shows a number of unique biocatalytic properties among hydrolases, e.g. high thermostability and stability at acidic pH ranges and the acceptance of tertiary and sterically hindered alcohols (de Maria, Carboni-Oerlemans et al. 2005). CALA also has a low sequence similarity to other members of the α/β hydrolase fold including lipase B. Therefore it was not included in previous versions of the LED. Only after its structure was recently determined (Ericsson, Kasrayan et al. 2008), a detailed analysis of its structure identified CALA unambiguously as a member of the α/β hydrolase family. However, in this structure the active site is not accessible to a substrate, therefore the molecular details of substrate binding or the existence of a possible lid are still elusive.

8.3.3. Results

Database content and layout

Release 3.0 of the Lipase Engineering Database (LED) contains 18582 proteins with 24782 sequence and 656 structure entries of which about 14000 protein and 540 structure entries are new. Six new homologous families and one new superfamily (the "Candida antarctica lipase A like" superfamily) have been added to the LED in the update process. Seed sequences for the new "Candida antarctica lipase A like" superfamily (LED identifier: abH38.01) included the sequence from the resolved crystal structure and three sequences of homologous lipases from other organisms which showed high sequence similarity to *Candida antarctica* lipase A (CALA) (Ericsson, Kasrayan et al. 2008) (Tab.1). The four largest superfamilies in release 3.0 contain 50% of all proteins in the LED: The "Cytosolic Hydrolases" superfamily (LED identifier: abh08) with 3188 proteins, containing epoxide hydrolases and haloalkane dehalogenases, the "Carboxylesterases" superfamily (LED identifier: abh01) with 2998 proteins, containing a wide range of carboxylesterases, such as acetylcholine esterases and bile salt activated lipases, the "Moraxella lipase 2 like" superfamily (LED identifier: abh04) with 1781 proteins containing mainly lipases and carboxylesterases, and the "Microsomal Hydrolases" superfamily (LED identifier: abh09) with 1336 proteins, containing microsomal epoxide hydrolases and peptidases. The "Cytosolic Hydrolases" and "Microsomal Hydrolases" superfamilies (abh08 and abh09) belong to the GX-class of α/β hydrolases, the "Carboxylesterases" and "Moraxella lipase 2 like" superfamilies (abh01 and abh04) belong to the GGGX-class of α/β hydrolases.

Candida antarctica lipase A protein family

The "Candida antarctica lipase A like" superfamily contains one crystal structure and 39 sequences, assigned to 32 proteins. They were grouped into four homologous families based on sequence similarity (Figure 1): The "Candida antarctica lipase A like" homologous family consisting of Lipase A from *C. antarctica*, the "Malassezia lipase like" homologous family consisting entirely of lipases and esterases from *Malassezia globosa* or *Malassezia furfur*, the "Candida albicans lipase like" homologous family consisting of various isoforms of the secretory lipase from *Candida albicans*, and the "Aspergillus lipase like" homologous family consisting mainly of hypothetical or putative lipases, mostly from *Aspergillus.* All 32 proteins are from organisms belonging to the subkingdom Dikarya of the kingdom Fungi. 12 proteins are classified as either lipases or esterases in GenBank (Benson, Karsch-Mizrachi et al. 2009)

while 20 proteins are classified as putative or hypothetical. The only structure entry in this superfamily is from the recently resolved crystal structure of CALA (Ericsson, Kasrayan et al. 2008). Based on the structure of the oxyanion hole, CALA can be classified as a Y-class lipase, and Tyr 93 was identified as the oxyanion hole forming amino acid. A structural comparison with other structures from the LED identified a structure from the deacetylase superfamily (LED identifier: abH26) as most closely related. A detailed structural alignment of CALA (PDB: 2VEO) with the structure of the *Bacillus subtilis* deacetylase (PDB: 1L7A) from the deacetylase superfamily showed a superimposition of the common α/β hydrolase fold including the catalytic triad (2VEO: Ser184, Asp334, His366; 1L7A: Ser181, Asp269, His298), despite having a low overall sequence identity of only 15%. Structural differences between the two structures are found in the cap and the C-terminal region. The cap region, located between β-strands 6 and 7, often confers substrate specificity or additional functions to the enzyme (Wei, Contreras et al. 1999). In the case of CALA, the cap region is involved in forming the tunnel like binding site for the acyl moiety (Ericsson, Kasrayan et al. 2008). For *B. subtilis* deacetylase, the cap region partially shields the active site from the solvent (Vincent, Charnock et al. 2003). The cap region of CALA consists of six α-helices, while the cap region of *B. subtilis* deacetylase consists of only four α-helices (Figure 2). Three α-helices in both proteins are found at identical positions. CALA shows an insert of three additional α-helices after the first two conserved α-helices and is missing the last α-helix of the cap present in *B. subtilis* deacetylase (Figure 2). In order to identify residues in the cap region which are conserved between the "Candida antarctica lipase A like" and "Deacetylases" superfamilies, a multisequence alignment of each family was performed. The two family alignments were aligned using a structural alignment of the two protein structures (Figure 3). The alignment demonstrated that despite the high structural similarity, there are no conserved residues in the cap region of both protein families. Two hidden Markov models of the three inserted α-helices of CALA and the α-helices shared by both proteins were created and used to search against all other protein families of the LED. No sequences with a significant similarity were found in the entire database, demonstrating that the sequence of the cap region of the "Candida antarctica lipase A like" superfamily is unique.

In comparison to the *B. subtilis* deacetylase, CALA has two additional β–strands (9 and 10) in the C-terminal region (Figure 2). They are positioned directly above the active site and prevent a direct access of the substrate to the active site. We assume that the β–strands 9 and 10 perform a lid like function for CALA since movement of the two β-strands would allow substrate access to the active site of CALA from a similar direction as for the *B. subtilis*

deacetylases (Vincent, Charnock et al. 2003). A comparison of the substrate binding sites of both proteins showed that the alcohol binding site is similar in both proteins and provides ample space for alcohol moieties of substrates (Figure 4). Therefore, both proteins are expected to accept a variety of bulky alcohols. The binding sites for the acyl moieties are highly different. CALA has a long, tunnel like binding site, while the *B. subtilis* deacetylase has a small cavity which is part of a cleft on the protein surface. Therefore, the acyl moieties of the substrates are expected to differ significantly between both enzymes. CALA is expected to accept mediumt to long chain fatty acids, while the *B. subtilis* deacetylase is limited to short-chain acyl moieties. Thus, despite the overall similarity between CALA and the *B. subtilis* deacetylase, the acyl binding site is fundamentally different.

However, the binding site of CALA shows surprising similarity to another lipase, *Candida rugosa* lipase (CRL). For CRL, two different structural confirmations have been resolved, an open conformation (1CRL) (Grochulski, Li et al. 1993), and a closed conformation (1TRH) (Grochulski, Li et al. 1994) where the lid of CRL is blocking the substrate access to the active site. CRL has a cap region between β-strands 6 and 7, consisting of four α-helices (Figure 2). The substrate binding site of CRL consists of a long tunnel for the acyl moiety of the substrate and provides ample space for the alcohol moiety of the substrate (Figure 4). Despite having a lower overall structure similarity to CALA than the *B. subtilis* deacetylase, the binding sites of CALA and CRL are highly similar (Figure 4). Both provide space for large, bulky alcohol moieties of the substrate and have a tunnel like binding site for the acyl moiety. Both proteins posses a lid which covers the active site and prevents direct access to the substrate binding site in its closed state. The lid of CRL lipase is formed by a α-helix between β-strands 1 and 2 and is located in the N-terminal region while the putative lid in CALA is formed by the two C-terminal β-strands 9 and 10 (Figure 2).

8.3.4. Discussion

The LED contains annotated and systematically classified protein families of α/β hydrolases. It has been shown to be a useful tool for the systematic analysis of protein families. Previous work employed the LED and BLAST in order to identify novel enzymes belonging to the α/β hydrolase fold (Lammle, Zipper et al. 2007; Kim, Oh et al. 2009). A model for the prediction of protein solubility was developed and refined by performing a comprehensive analysis of the protein families of the LED (Koschorreck, Fischer et al. 2005). A further study involved the systematic analysis of protein families of the LED in regard to the distribution and conservation of functionally relevant rare codons (Widmann, Clairo et al. 2008).

Since the first release of the LED (Pleiss, Fischer et al. 2000), more than 14000 new α/β hydrolases became available and were integrated in the release 3.0. As a case study for the utility of the highly enriched and annotated database, the newly introduced superfamily of CALA was analysed and compared to other protein structures in the LED. The goal was to characterise the sequence and structure of CALA in comparison to other α/β hydrolases despite its low sequence similarity and to understand the molecular basis of substrate recognition.

While CALA shows structural similarity to the deacetylase family, the substrate specificity of both enzymes differs, which is consistent with the differences observed in the substrate binding sites of both proteins. In contrast, the lipase from *C. rugosa* (CRL), which shows a lower overall structural similarity to CALA, is remarkably similar in regard to the substrate binding site. The structural similarities and differences are in accordance with experimentally observed substrate specificities of the three enzymes. All three proteins have a spacious alcohol binding site. The *B. subtilis* deacetylase accepts a wide variety of bulky substrates like cephalosporin C and xylose (Vincent, Charnock et al. 2003). CALA and CRL also accept bulky substrates, ranging from primary alcohols to sterically hindered secondary alcohols and even tertiary alcohols (Kirk and Christensen 2002; Akoh, Lee et al. 2004).

The tunnel like binding site of CALA allows the enzyme to accept esters of long chain fatty acids (Kirk and Christensen 2002; Pfeffer, Richter et al. 2006). The similar tunnel like acyl binding site of CRL also accepts fatty acids up to a chain length of 18 (Pfeffer, Richter et al. 2006). In contrast, the small acyl binding site of the *B. subtilis* deacetylase is unable to accept large acyl groups and is restricted towards acetyl moieties (Vincent, Charnock et al. 2003). Experimentally, CALA and CRL have been shown to display interfacial activation (Grochulski, Li et al. 1993; Martinelle, Holmquist et al. 1995). While a lid in CRL has been

localized and the open and closed form of CRL has been crystallized (Grochulski, Li et al. 1993; Grochulski, Li et al. 1994), the lid function of the β-strands 9-10 in CALA remains to be experimentally verified. However, the similarities to CRL suggest a substrate access involving the movement of β-strands 9-10.

8.3.5. Conclusions

The analysis of the newly introduced protein family of *Candida antarctica* lipase A demonstrates the strength of our database approach by providing a large set of protein families which share a common protein fold despite an overall low sequence similarity. By combining both, structural and sequential information of a large number of proteins a thorough analysis and classification of proteins of interest is made possible.

8.3.6. Availability and Requirements

The Lipase Engineering Database (LED) is online accessible at http://www.led.uni-stuttgart.de. All information on families of sequence and structure data, as well as alignments, phylogenetic trees, and family-specific profiles can be accessed by manual download.

8.3.7. Methods

Structural comparisons and alignments

Comparison of structures were carried out using DALI (Holm and Sander 1995). The structure of CALA was compared against 28 representative structures from all superfamilies. To identify the most closely related superfamilies, only structures which could be aligned to more than 50% of the residues of CALA were considered. Structural alignments of proteins were performed by STAMP (Russell and Barton 1992).

For superfamilies which share a close structural relationship but a low overall sequence identity, a two step strategy was used in order to obtain a more significant multisequence alignment. First, a multisequence alignment for each of the two superfamilies was carried out separately. Then a structural alignment, between reference structures from each protein family was performed using STAMP (Russell and Barton 1992). The multisequence alignments where then aligned against the structure from their respective protein family.

Sequence analysis

Multisequence alignments for all protein families were generated using ClustalW (Thompson, Higgins et al. 1994) with a gap opening and extension penalties of 10 and 0.2, respectively. Hidden Markov models were created using HMMER (Eddy 1998).

Database model

The implemented data model is based on Firebird (Firebird) and is based on the previously published (Fischer, Thai et al. 2006) data model (Supplementary File 1, Figure S1). Protein families are organised on the level of homologous families and superfamilies based on their sequence similarity. The database is updated by an automated Perl (PERL) script. It performs a BLAST (Altschul, Gish et al. 1990) search against the current version of the non-redundant sequence database an NCBI (Benson, Karsch-Mizrachi et al. 2009) for each sequence entry with an E-value cut-off of 10^{-50}. Crystal structure information referring to new sequence entries is updated as well. New sequence and structure entries are assigned to homologous families and superfamilies based on sequence similarity. New families which consisted of only one putative protein entry where not included. Annotation information of residues is either taken directly from the according GenBank entry or is transferred to new sequences using the DWARF graphical user interface. Annotation information is then transferred to the newly integrated sequences.

8.3.8. Authors' contributions

MW performed all analyses regarding CALA and drafted the manuscript. MW and BJ performed the update of the database and the manual curation process. JP supervised the project and finalized the manuscript.

8.3.9. Acknowledgements

We acknowledge the valuable contribution of Robert Radloff for help in the annotation process and of Florian Wagner for the programming of the dynamic user interface. The work was carried out in the framework of the IP-project 'Sustainable Microbial and Biocatalytic Production of Advanced Functional Materials' (BIOPRODUCTION / NMP-2-CT-2007-026515) funded by the European Commission.

8.3.10. References

Akoh, C. C., G. C. Lee, et al. (2004). "Protein engineering and applications of Candida rugosa lipase isoforms." Lipids 39(6): 513-526.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol 215(3): 403-10.

Barth, S., M. Fischer, et al. (2004). "The database of epoxide hydrolases and haloalkane dehalogenases: one structure, many functions." Bioinformatics 20(16): 2845-2847.

Benson, D. A., I. Karsch-Mizrachi, et al. (2009). "GenBank." Nucleic Acids Res 37(Database issue): D26-31.

de Maria, P. D., C. Carboni-Oerlemans, et al. (2005). "Biotechnological applications of Candida antarctica lipase A: State-of-the-art." Journal of Molecular Catalysis B-Enzymatic 37(1-6): 36-46.

Eddy, S. (1998). "HMMER." from http://hmmer.wustl.edu/.

Ericsson, D. J., A. Kasrayan, et al. (2008). "X-ray structure of Candida antarctica lipase a shows A novel lid structure and a likely mode of interfacial activation." Journal of Molecular Biology 376(1): 109-119.

Firebird. "Firebird." from http://sourceforge.net/projects/firebird.

Fischer, M. and J. Pleiss (2003). "The Lipase Engineering Database: a navigation and analysis tool for protein families." Nucleic Acids Res 31(1): 319-21.

Fischer, M., Q. K. Thai, et al. (2006). "DWARF--a data warehouse system for analyzing protein families." BMC Bioinformatics 7: 495.

Gotor-Fernandez, V., E. Busto, et al. (2006). "Candida antarctica lipase B: An ideal biocatalyst for the preparation of nitrogenated organic compounds." Advanced Synthesis & Catalysis 348(7-8): 797-812.

Grochulski, P., Y. Li, et al. (1993). "Insights into interfacial activation from an open structure of Candida rugosa lipase." J Biol Chem 268(17): 12843-7.

Grochulski, P., Y. Li, et al. (1994). "Two conformational states of Candida rugosa lipase." Protein Sci 3(1): 82-91.

Grochulski, P., Y. G. Li, et al. (1993). "Insights into Interfacial Activation from an Open Structure of Candida-Rugosa Lipase." Journal of Biological Chemistry 268(17): 12843-12847.

Holm, L. and C. Sander (1995). "Dali: a network tool for protein structure comparison." Trends Biochem Sci 20(11): 478-80.

Kim, E. Y., K. H. Oh, et al. (2009). "Novel cold-adapted alkaline lipase from an intertidal flat metagenome and proposal for a new family of bacterial lipases." Appl Environ Microbiol 75(1): 257-60.

Kirk, O. and M. W. Christensen (2002). "Lipases from Candida antarctica: Unique Biocatalysts from a Unique Origin." Org. Proc. Res. 6(4): 446–451.

Koschorreck, M., M. Fischer, et al. (2005). "How to find soluble proteins: a comprehensive analysis of alpha/beta hydrolases for recombinant expression in E. coli." BMC Genomics 6(1): 49.

Lammle, K., H. Zipper, et al. (2007). "Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning." J Biotechnol 127(4): 575-92.

Martinelle, M., M. Holmquist, et al. (1995). "On the interfacial activation of Candida antarctica lipase A and B as compared with Humicola lanuginosa lipase." Biochim Biophys Acta 1258(3): 272-6.

Ollis, D. L., E. Cheah, et al. (1992). "The alpha/beta hydrolase fold." Protein Eng 5(3): 197-211.

Orrenius, C., T. Norin, et al. (1995). "The Candida antarctica lipase B catalysed kinetic resolution of seudenol in non-aqueous media of controlled water activity." Tetrahedron-Asymmetry 6(12): 3023-3030.

Orrenius, C., N. Ohrner, et al. (1995). "Candida-Antarctica Lipase-B Catalyzed Kinetic Resolutions - Substrate Structure Requirements for the Preparation of Enantiomerically Enriched Secondary Alcanols." Tetrahedron-Asymmetry 6(5): 1217-1220.

PERL. "PERL." from http://www.perl.org/.

Pfeffer, J., S. Richter, et al. (2006). "High yield expression of Lipase A from Candida antarctica in the methylotrophic yeast Pichia pastoris and its purification and characterisation." Applied Microbiology and Biotechnology 72(5): 931-938.

Pleiss, J., M. Fischer, et al. (2000). "Lipase engineering database - Understanding and exploiting sequence-structure-function relationships." Journal of Molecular Catalysis B-Enzymatic 10(5): 491-508.

Russell, R. B. and G. J. Barton (1992). "Multiple Protein-Sequence Alignment from Tertiary Structure Comparison - Assignment of Global and Residue Confidence Levels." Proteins-Structure Function and Genetics 14(2): 309-323.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res 22(22): 4673-80.

Vincent, F., S. J. Charnock, et al. (2003). "Multifunctional xylooligosaccharide/cephalosporin C deacetylase revealed by the hexameric structure of the Bacillus subtilis enzyme at 1.9A resolution." J Mol Biol 330(3): 593-606.

Wei, Y., J. A. Contreras, et al. (1999). "Crystal structure of brefeldin A esterase, a bacterial homolog of the mammalian hormone-sensitive lipase." Nat Struct Biol 6(4): 340-5.

Widmann, M., M. Clairo, et al. (2008). "Analysis of the distribution of functionally relevant rare codons." BMC Genomics 9:207

8.3.11. Figures

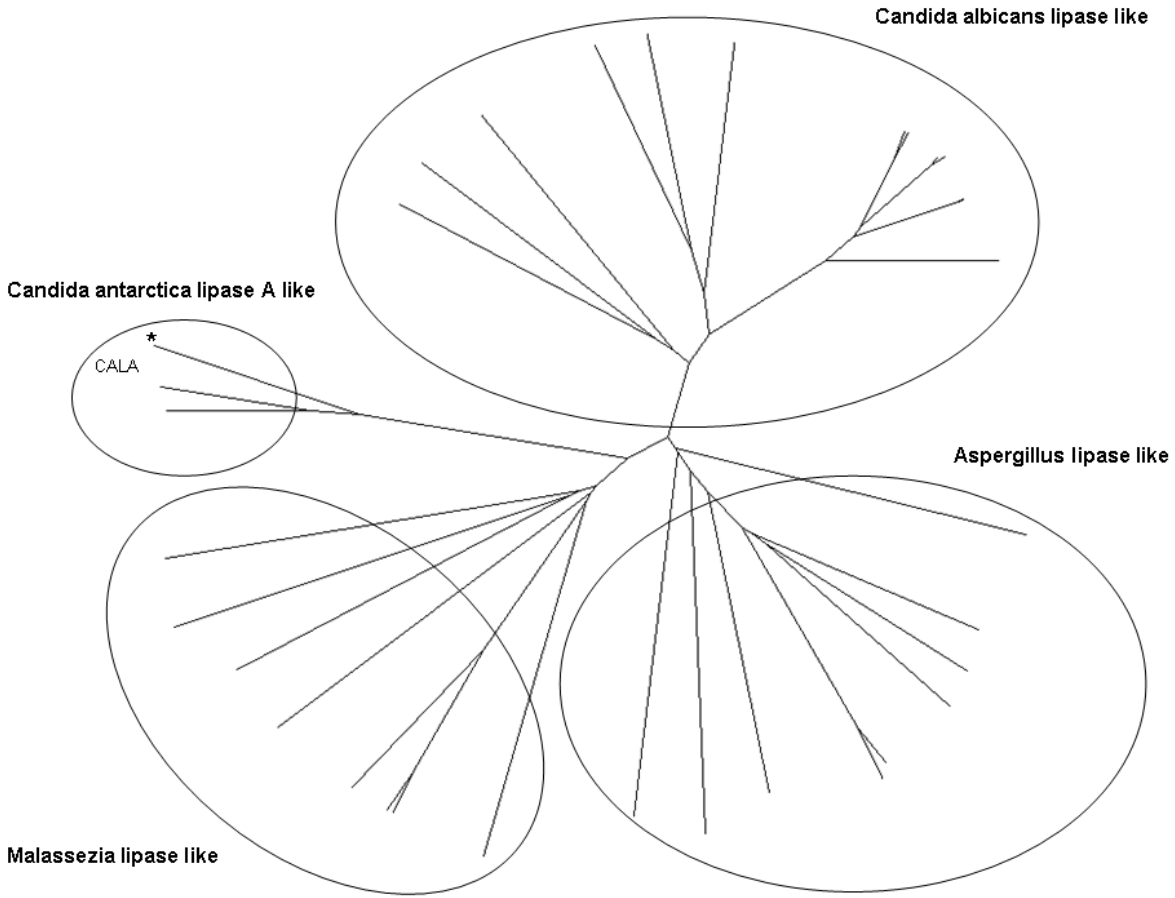


Figure 1 – Phylogenetic tree of the "Candida antarctica lipase A like" superfamily

The superfamily consists of 4 homologous families based on sequence similarity. The sequence of *C. antarctica* lipase A is indicated.

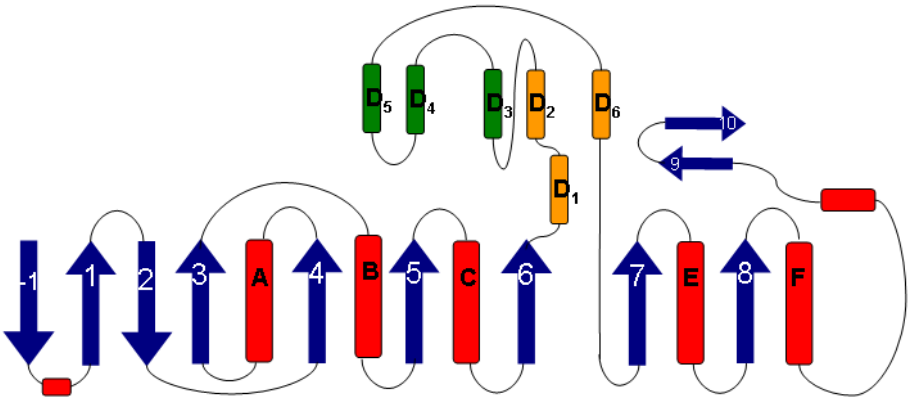
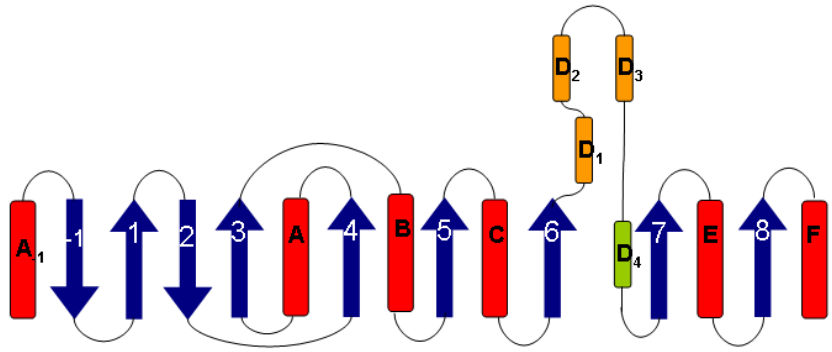
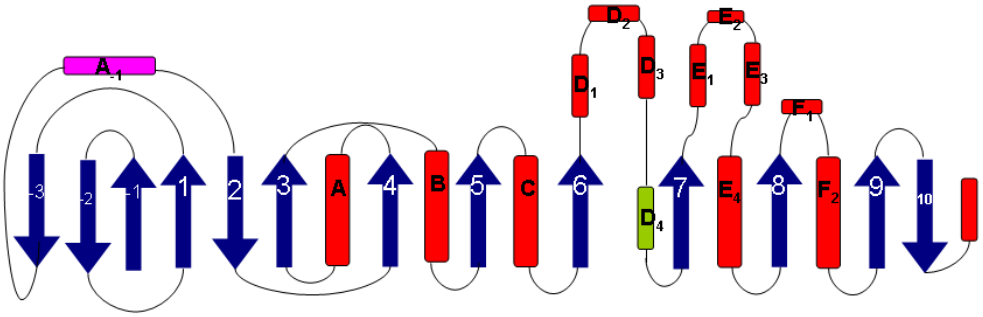
Candida antarctica **lipase A (PDB: 2VEO)**	
Bacillus subtilis **deacetylase (PDB: 1L7A)**	
Candida rugosa **lipase (PDB: 1CRL)**	

Figure 2 – Topology diagrams of *Candida antarctica* lipase A, *Bacillus subtilis* deacetylase and *Candida rugosa* lipase

The shared cap region between *C. antarctica* lipase A and *B. subtilis* deacetylase is colored orange. The additional 3 α-helices of the cap region in CALA are labelled D_3, D_4 and D_5 and colored in dark green. The C-terminal, presumably lid forming β-strands of CALA are labelled 9 and 10. The lid forming α-helix of *C. rugosa* lipase is labelled A_{-1}.

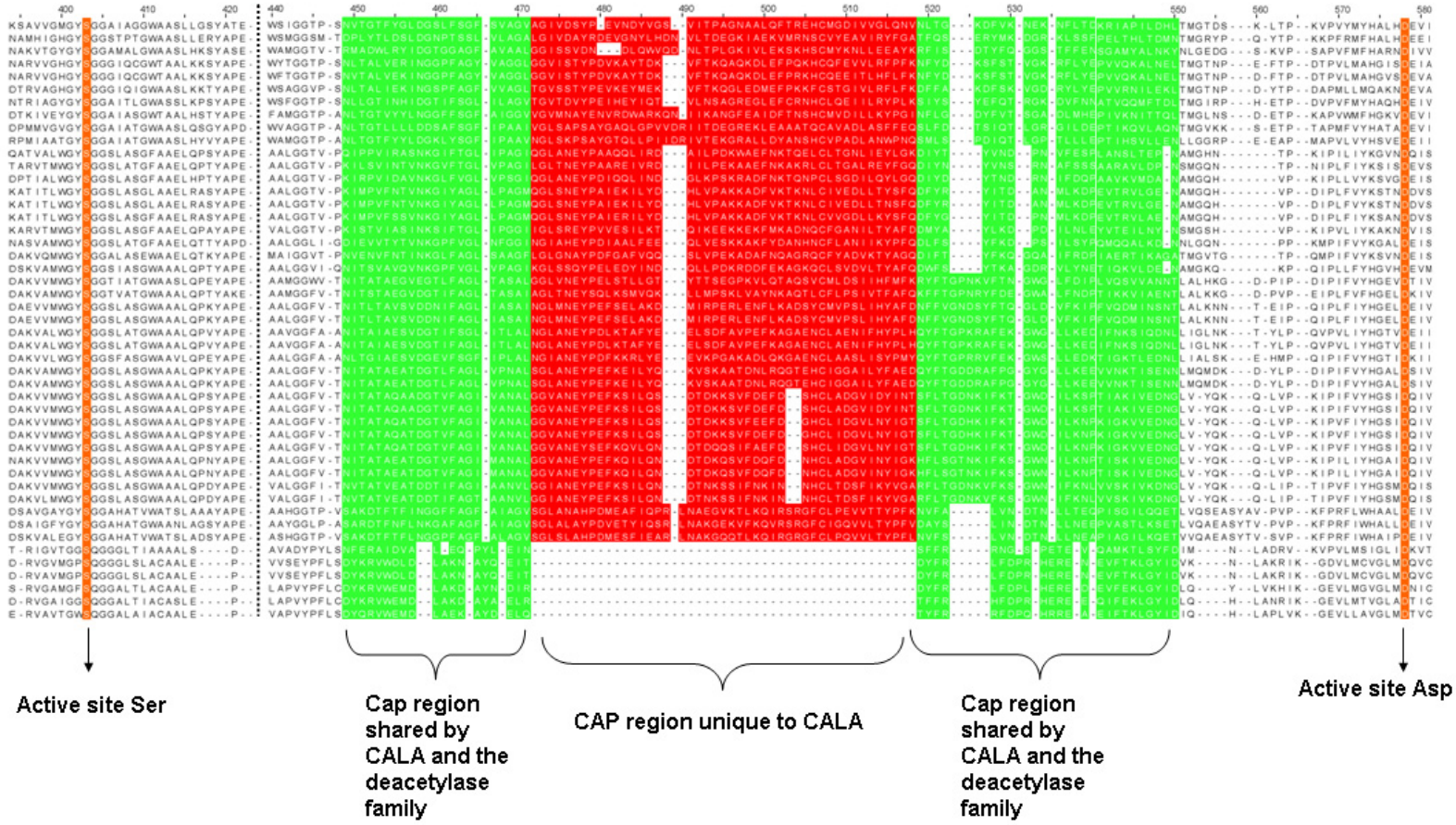


Figure 3 – Multisequence alignment between the "Candida antarctica lipase A ike" and "Deacetylases" superfamilies

Shown is an excerpt of the alignment containing the active site Ser and Asp and the cap region. The cap region of both superfamilies is colour coded: shared regions green; the additional 3 α–helices of the "Candida antarctica lipase A like" superfamily red. Columns containing active site residues are coded orange.

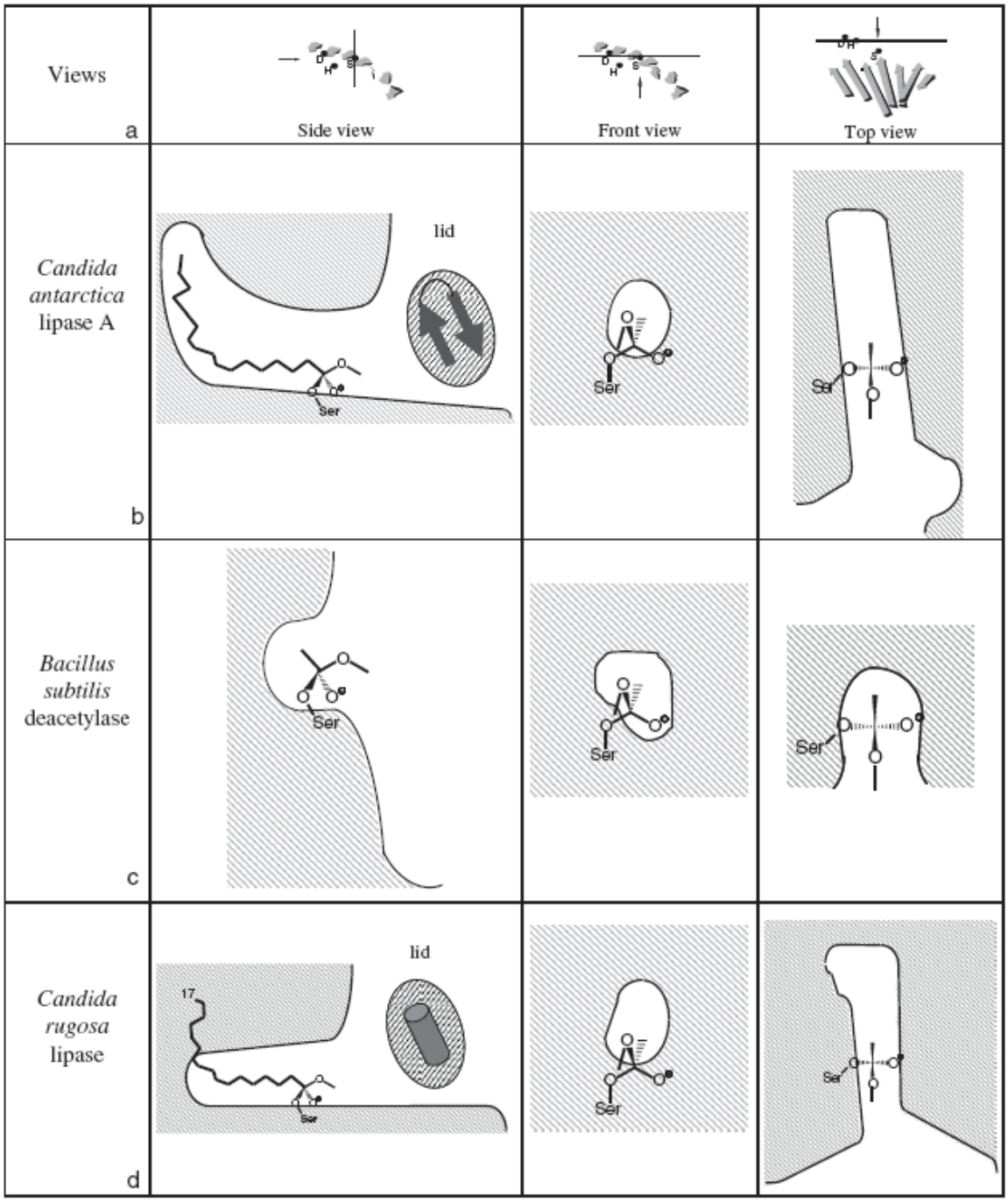


Figure 4 - Shape of the binding site of *Candida antarctica* lipase A, *Bacillus subtilis* deacetylase and *Candida rugosa* lipase

(a) Orientation of the cross-sections which are planes perpendicular to the paper plane and indicated by a straight line. The direction of the view is indicated by an arrow. Shape of the binding sites is displayed in side, front and top view for (b) *Candida antarctica* lipase A, (c) *Bacillus subtilis* deacetylase, and (d) *Candida rugosa* lipase. A model of the acyl moiety of the substrate is displayed, the alcohol moiety is not shown for clarity. The position of the lid in a closed state for *Candida antarctica* lipase A and *Candida rugosa* lipase is indicated.

8.3.12. Tables

Table 1. Seed sequences for the "Candida antarctica lipase A like" protein family.

Accession number (gi)	Organism	Homologous Family
160286179	*Candida antarctica*	Candida antarctica lipase A like
20429169	*Kurtzmanomyces sp. I-11*	Candida antarctica lipase A like
73765555	*Malassezia furfur*	Malassezia lipase like
71018653	*Ustilago maydis 521*	Candida antarctica lipase A like

8.3.13. Supplementary material

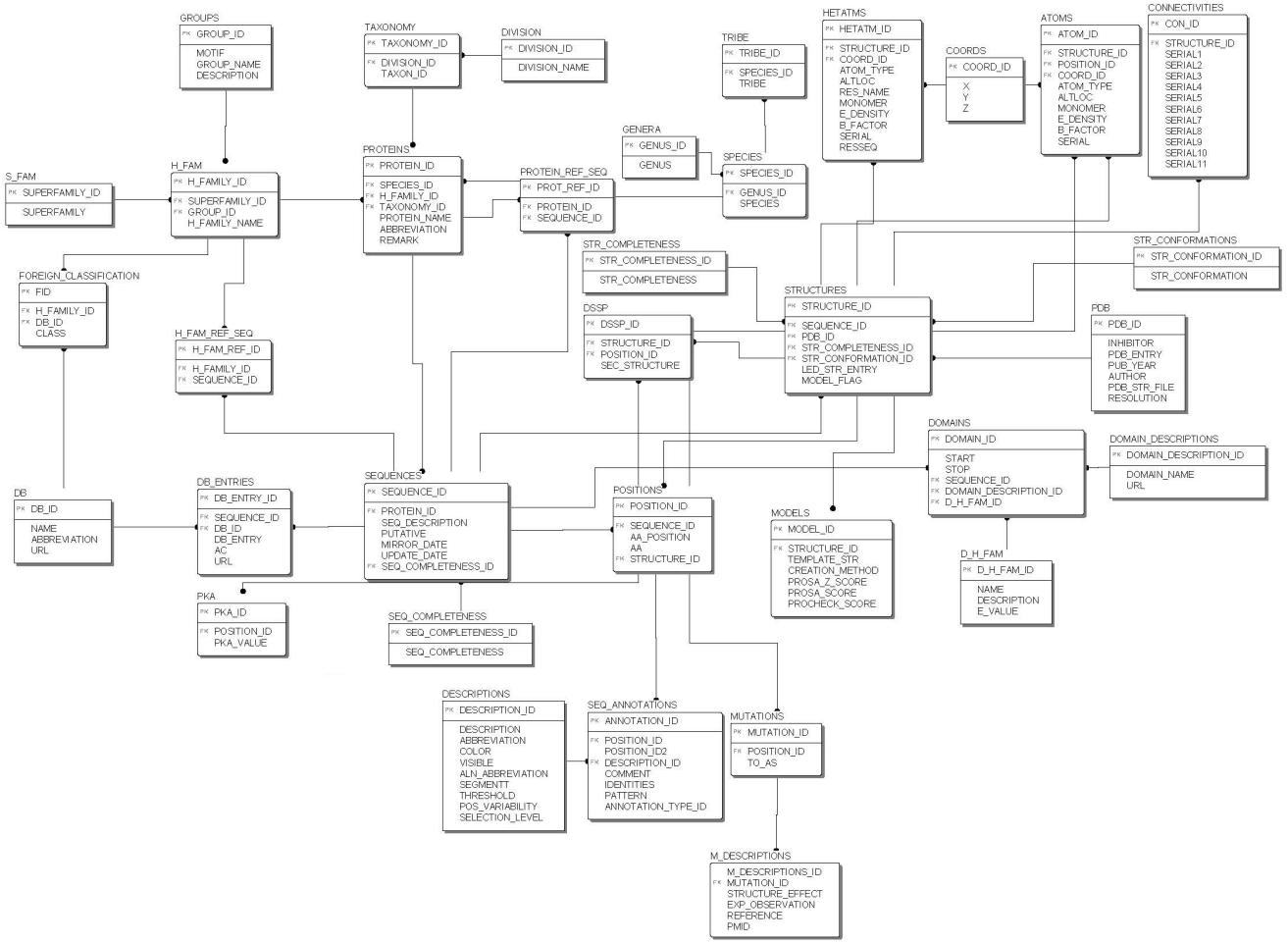


Figure S1 - Conceptual data scheme for the LED using Logical Data Structure (LDS) notation

Each database table is represented by a separate table. Primary key attributes are displayed in the header of the respective table.

8.4. The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations

Michael Widmann,[1] Robert Radloff,[1] and Jürgen Pleiss [1§]

[1] Institute of Technical Biochemistry, University of Stuttgart,
Allmandring 31, 70569 Stuttgart, Germany

Publikation erschienen bei *BMC Biochemistry* 2010, 11:9

8.4.1. Abstract

Background

Thiamine diphosphate (ThDP)-dependent enzymes form a vast and diverse class of proteins, catalyzing a wide variety of enzymatic reactions including the formation or cleavage of carbon-sulfur, carbon-oxygen, carbon-nitrogen, and especially carbon-carbon bonds. Although very diverse in sequence and domain organisation, they share two common protein domains, the pyrophosphate (PP) and the pyrimidine (PYR) domain. For the comprehensive and systematic comparison of protein sequences and structures the Thiamine diphosphate (ThDP)-dependent Enzyme Engineering Database (TEED) was established.

Results

The TEED (http://www.teed.uni-stuttgart.de) contains 12048 sequence entries which were assigned to 9443 different proteins and 379 structure entries. Proteins were assigned to 8 different superfamilies and 63 homologous protein families. For each family, the TEED offers multisequence alignments, phylogenetic trees, and family-specific HMM profiles. The conserved pyrophosphate (PP) and pyrimidine (PYR) domains have been annotated, which allows the analysis of sequence similarities for a broad variety of proteins. Human ThDP-dependent enzymes are known to be involved in many diseases. 20 different proteins and over 40 single nucleotide polymorphisms (SNPs) of human ThDP-dependent enzymes were identified in the TEED.

Conclusion

The online accessible version of the TEED has been designed to serve as a navigation and analysis tool for the large and diverse family of ThDP-dependent enzymes.

8.4.2. Background

Since the discovery of the first thiamine diphosphate (ThDP)-dependent enzyme in 1937, a multitude of them has been described and their catalytic mechanism was intensively analysed (Schellenberger 1998; Jordan 2003; Frank, Leeper et al. 2007). ThDP-dependent enzymes catalyze a wide variety of enzymatic reactions and therefore were assigned to the families of oxidoreductases, transferases, or lyases (Bairoch, Bougueleret et al. 2008). The formation or cleavage of carbon-sulfur, carbon-oxygen, carbon-nitrogen, and especially carbon-carbon bonds are of utmost interest for bioorganic synthesis and organocatalysis (Zeitler 2005; Enders, Niemeier et al. 2007). Because of their ability to form asymmetric C-C bonds, ThDP-dependent enzymes are versatile catalysts for a variety of biotransformations (Iding, Siegert et al. 1998; Pohl, Sprenger et al. 2004; Stillger, Pohl et al. 2006; Berthold, Gocke et al. 2007; Demir, Ayhan et al. 2007; Mueller, Gocke et al. 2009). In addition, the ThDP-dependent enzyme family has been shown to possess a wide substrate spectrum ranging from small compounds like formaldehyde to bulky hydroxyl-phytanoyl-CoA molecules (Bornemann, Crout et al. 1993; Casteels, Foulon et al. 2003). For pharmacology, ThDP-dependent enzymes of human origin are of special interest. They have been identified as being involved in a variety of diseases like Alzheimer's disease and diabetes (Shils 2006), and also play a role in tumor proliferation (Zhao and Zhong 2009). Their highly diverse substrate specificity and catalytic activity is reflected in their sequence and structure which differs significantly between different families of ThDP-dependent enzymes. During the course of evolution, shuffling, rearrangement, and fusion of domains, as well as mutation, and gene duplications have led to the enormous diversity of ThDP-dependent enzymes (Duggleby 2006; Costelloe, Ward et al. 2008). However, all ThDP-dependent enzymes contain at least two conserved domains, the pyrophosphate (PP) and the pyrimidine (PYR) domain, which have a similar structure (Duggleby 2006) and are essential for binding and activating ThDP (Wang, Martin et al. 1997). The PYR domain has a conserved catalytic glutamic acid while the PP domain contains a conserved $GDX_{25\text{-}30}N$ motif (Hawkins, Borges et al. 1989; Candy and Duggleby 1998; Fang, Nixon et al. 1998; Costelloe, Ward et al. 2008). In addition to these two domains, additional domains were found such as the the transhydrogenase dIII domain (TH3) and the transketolase C-terminal domain (TKC) (Cromartie and Walsh 1976; Duggleby 2006; Costelloe, Ward et al. 2008). These additional domains are often not well characterised and in many cases their function in the catalytic process remains obscure (Costelloe, Ward et al.

2008). A unified classification scheme for ThDP-dependent enzymes based on a comprehensive analysis of sequence and structure does not yet exist. Based on a structural comparison, it was suggested that a total of 4 families should be sufficient to describe ThDP-dependent enzymes: DC (decarboxylases), TK (transketolases), OR (oxidoreductases), and KD (2-ketoacid dehydrogenase) (Duggleby 2006). A sequence based evolutionary analysis suggested at least 6 different families, namely TK (transketolases)-like, PFRD (pyruvate ferredoxin reductase), 2OXO (2-oxoisovalerate dehydrogenase)-like, PDC (pyruvate decarboxylase)-like, SPDC (sulfopyruvate decarboxylase), and PPDC (phosphopyruvate decarboxylase) (Costelloe, Ward et al. 2008).

We established the Thiamine diphosphate dependent Enzyme Engineering Database (TEED) as a tool for a comprehensive and systematic comparison of ThDP-dependent enzymes from different protein families and annotated the conserved PP- and PYR domains. Thus, the TEED is the first data resource of ThDP-dependent enzymes which combines information on the individual protein families, sequence alignments and a consistent annotation of the conserved PYR and PP domains.

8.4.3. Construction and Content

Source Data

The Thiamine diphosphate (ThDP)-dependent Enzyme Engineering Database (TEED) was established by utilising the data warehouse system DWARF (Fischer, Thai et al. 2006). The DWARF system is a collection of tools for the automated retrieval and integration of protein sequences and structures from different source databases and their subsequent integration into a local data warehouse system. The initial step in the construction of the database consisted of the selection of seed sequences of 62 proteins which represent members of the different ThDP-dependent protein families (Tab. S1, supplementary file 1). Seed sequences were selected based on the enzymatic activity of the protein and the structural arrangement of protein domains. This selection was based on previous work (Duggleby 2006; Costelloe, Ward et al. 2008) which divided the members of the ThDP-dependent enzymes in different protein families.

Database establishment

The combination of previous classification schemes resulted in 8 different superfamilies, DC (decarboxylase), TK (transketolase), OR (oxidoreductase), and two subfamilies K1 and K2 of the KD (2-ketoacid dehydrogenase) family. In addition to these families, the SPDC (sulfopyruvate decarboxylase), the PPDC (phosphopyruvate decarboxylase), and the KDH (α-ketoglutarate dehydrogenase) family were included (Figure 1). To populate the TEED, a BLAST search against the sequence database at NCBI (http://www.ncbi.nlm.nih.gov) was carried out for each seed sequence with an E-value cut off of 10^{-5}. New protein entries were assigned to a homologous protein family based on their sequence similarity to one of the seed sequences. If the sequence similarity was less than 60%, a protein was assigned to a new homologous family. The families were subsequently manually evaluated and adjusted: protein fragments were merged into the respective homologous family, and proteins with high sequence similarity but a different domain organization were separated into different protein families. This resulted in 63 different homologous families.

Sequence entries with more than 98% sequence identity which shared the same source organism were assigned to the same protein entry. If more than one sequence was assigned to the same protein, the longest sequence was set as the reference sequence. If structural information was available for protein entries, structural monomers were downloaded from the

Protein Data Bank (Berman, Battistuz et al. 2002) and stored as structure entries. Secondary structure information was calculated by DSSP (Kabsch and Sander 1983) and displayed in the annotated multisequence alignments which were generated by ClustalW (v1.83) with default parameters (Thompson, Higgins et al. 1994). Additional annotation on structurally or functionally relevant residues (active site, disulfide bridges, signal peptide) were extracted from the NCBI entry and the respective residues were annotated in the TEED. Abbreviations for the established protein families are available in tabular form (Tab. S1, supplementary file 1).

Features and functionalities

The online version of the TEED offers pre-calculated multisequence alignments and can be browsed by families, organisms, or structures. Phylogenetic trees are visualized by the program PHYLODENDRON (PHYLODENDRON). The PP and PYR domain of each ThDP-dependent protein family was manually annotated. If structural information for a protein homologous family was available, a structural alignment of the available structures was performed using STAMP (Russell and Barton 1992).

If no structure information was available, a set of sequences from the respective homologous protein family was selected and used to create a multisequence alignment. A reliable alignment was ensured by performing this analysis for each homologous family separately to ensure a high degree of sequence similarity. This set consisted of full length sequences from different organisms, excluding protein fragments. The information on the domain boundaries for these sequences was retrieved from InterProScan (Zdobnov and Apweiler 2001). Since in many cases, information on the exact N- and C- terminal boundaries for each domain was inconsistent, the boundaries were preferably assigned in well conserved regions rather than in more variable regions. For each multisequence alignment, a Hidden Markov Model (HMM) was created using HMMER (Eddy 1998). For each homologous family, the individual HMM was used to perform alignments of every protein sequence of this family against the annotated multisequence alignment. Based on this alignment, the PP and PYR domain annotations from the annotated sequences were transferred to every sequence (Figure 2). Annotation information of the PP and PYR domains is displayed for each pre-calculated alignment of homologous families or superfamilies and allows the systematic analysis and evaluation of properties and relationships of these domains. The TEED consists of 12048 sequence entries which were assigned to 9443 different proteins and 379 structure entries. The largest superfamily is the DC family. It consists of more than 4000 sequence entries and accounts for

35% of all sequence entries. The TK and OR families are of comparable size (2600 and 2257 sequence entries, respectively) and account for 21% and 19%, respectively. The source organism of the majority of ThDP-dependent enzymes in the TEED are bacteria (87%).

Human ThDP-dependent enzymes

66 sequence entries from the TEED are of human origin (excluding sequences from crystal structure chains). Due to their medical importance they were systematically analysed. All human ThDP-dependent enzymes belong to only three superfamilies, the DC, TK, and K2 superfamily (Tab. 1). The 66 sequences belong to 20 different proteins with several isoforms. The transketolase (gi: 205277463) with most isoforms (12) is implicated in the latent genetic disease Wernicke-Korsakoff syndrome (Wang, Martin et al. 1997) and has been found to be differentially expressed in the dorsolateral prefrontal cortex from patients with schizophrenia. Another human ThDP-dependent enzyme with many isoforms (7) is the 2-oxoisovalerate dehydrogenase subunit alpha (gi: 548403), also known as branched-chain alpha-keto acid dehydrogenase. This protein is involved in the catabolism of amino acids like isoleucine, leucine, and valine, and a defect causes the accumulation of these amino acids which leads to the maple syrup urine disease (Podebrad, Heil et al. 1999). One third of all sequence entries was labelled as ‘putative’ or ‘unnamed’ in the GenBank, and was assigned to a specific protein or protein family based on sequence similarity (Tab. 1). However, because the function and substrate specificity can vary considerably even between homologous proteins, the assignment of a biochemical property based on sequence similarity only should be regarded as putative. All sequence entries were compared to the respective full sequence and were subsequently classified as either fragments or SNPs. Fragments consist of parts of the full sequence but show no exchange of amino acids while SNPs always show an exchange of amino acids (Tab. 1).

8.4.4. Utility and discussion

The analysis of the humans ThDP-dependent enzymes led to a reliable classification of several, previously unclassified proteins and demonstrates the advantage of a highly enriched database of a specific protein family. SNPs have been shown to play an important role in tumor development (Mimori, Inoue et al. 2002; Martin, Broaddus et al. 2004) therefore a complete analysis for SNPs was included in the analysis of human ThDP-dependent enzymes.

This analysis of SNPs is limited to sequences retrieved from GenBank (Benson, Karsch-Mizrachi et al. 2009) and thus complements specialised SNP repositories such as the dbSNP (Sherry, Ward et al. 2001). Our analysis demonstrates that GenBank annotations are often incomplete and unreliable for the identification of proteins or protein variants. The transketolase (gi: 205277463) includes 12 different isoforms, of which 6 have been designated as protein fragments. Of these, only one sequence (gi: 193787540) shows an internal deletion, suggesting a truly altered protein product. The other 5 isoforms only show truncated N-termini and therefore could be sequencing artefacts of the original protein.

This kind of analysis is not limited to proteins from a specific organism but can be expanded to cover protein superfamilies or specific homologous families. It has been shown previously that a systematic classification of protein families can be used as a reliable framework for systematic analyses of protein families (Fischer and Pleiss 2003; Knoll, Hamm et al. 2009) and for the engineering of protein mutants with improved biochemical properties (Seifert and Pleiss 2009; Seifert, Vomund et al. 2009). With the implemented domain annotation, an analysis is not limited to the whole protein sequence but protein families can also be specifically analyzed for differences and conserved features in the PP and PYR domains.

Web accessibility

The database can be accessed on the level of sequence, structure, or organism. All protein entries link to the respective NCBI entries. Annotated multiple sequence alignments and phylogenetic trees are provided via the online accessible version of the TEED at http://www.teed.uni-stuttgart.de. For each family, the level of amino acid conservation is calculated by PLOTCON (Rice, Longden et al. 2000). BLAST searches (Altschul, Gish et al. 1990) can be performed against the TEED using a local BLAST interface. Updates for the TEED will be performed regularly using an automated scripting system. For new sequence entries referring to a new structure in the Protein Data Bank (PDB), structure information is updated as well. New sequence and structure entries are assigned to existing homologous families and superfamilies based on their sequence similarity.

8.4.5. Conclusions

The Thiamine diphosphate dependent Enzyme Engineering Database (TEED) has been designed to serve as a navigation and analysis tool for the large and diverse family of ThDP-dependent enzymes. The annotation of the conserved pyrophosphate (PP) and pyrimidine (PYR) domains allows for a direct comparison and analysis of these domains between different families. Thus the TEED is a valuable tool for the study of the protein families of ThDP-dependent enzymes.

8.4.6. Availability and requirements

The Thiamine diphosphate dependent Enzyme Engineering Database (TEED) is online accessible at http://www.teed.uni-stuttgart.de. All information on families, sequence and structure data, as well as alignments and phylogenetic trees can be accessed by manual download.

8.4.7. List of abbreviations

BLAST: Basic Local Alignment Search Tool
DSSP: Define Secondary Structure of Proteins
DWARF: Data warehouse system for analyzing protein families
HMM: Hidden Markov Model
SNP: Single-nucleotide polymorphism
TEED: Thiamine diphosphate dependent Enzyme Engineering Database
ThDP: Thiamine diphosphate

8.4.8. Authors' contributions

MW established and annotated the database and wrote the manuscript. RR assisted in the implementation of the database and contributed to writing of the manuscript. JP supervised the project and finalized the manuscript. All authors read and approved the final manuscript.

8.4.9. Acknowledgements

We acknowledge valuable contribution to the development of the domain annotation approach by Demet Sirim. We also thank Florian Wagner for support in the technical maintenance of the database. This work was supported by the DFG (PL145/6-1)

8.4.10. References

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Bairoch, A., L. Bougueleret, et al. (2008). "The Universal Protein Resource (UniProt)." Nucleic Acids Research **36**: D190-D195.

Benson, D. A., I. Karsch-Mizrachi, et al. (2009). "GenBank." Nucleic Acids Research **37**: D26-D31.

Berman, H. M., T. Battistuz, et al. (2002). "The Protein Data Bank." Acta Crystallogr D Biol Crystallogr **58**(Pt 6 No 1): 899-907.

Berthold, C. L., D. Gocke, et al. (2007). "Structure of the branched-chain keto acid decarboxylase (KdcA) from Lactococcus lactis provides insights into the structural basis for the chemoselective and enantioselective carboligation reaction." Acta Crystallographica Section D-Biological Crystallography **63**: 1217-1224.

Bornemann, S., D. H. G. Crout, et al. (1993). "Stereochemistry of the Formation of Lactaldehyde and Acetoin Produced by the Pyruvate Decarboxylases of Yeast (Saccharomyces Sp) and Zymomonas-Mobilis - Different Boltzmann Distributions between Bound Forms of the Electrophile, Acetaldehyde, in the 2 Enzymatic-Reactions." Journal of the Chemical Society-Perkin Transactions 1(3): 309-311.

Candy, J. M. and R. G. Duggleby (1998). "Structure and properties of pyruvate decarboxylase and site-directed mutagenesis of the Zymomonas mobilis enzyme." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1385**(2): 323-338.

Casteels, M., V. Foulon, et al. (2003). "Alpha-oxidation of 3-methyl-substituted fatty acids and its thiamine dependence." European Journal of Biochemistry **270**(8): 1619-1627.

Costelloe, S. J., J. M. Ward, et al. (2008). "Evolutionary analysis of the TPP-dependent enzyme family." J Mol Evol **66**(1): 36-49.

Cromartie, T. H. and C. T. Walsh (1976). "Escherichia-Coli Glyoxalate Carboligase - Properties and Reconstitution with 5-Deazafad and 1,5-Dihydrodeazafadh2." Journal of

Biological Chemistry **251**(2): 329-333.

Demir, A. S., P. Ayhan, et al. (2007). "Thiamine pyrophosphate dependent enzyme catalyzed reactions: Stereoselective C-Cbond formations in water." Clean-Soil Air Water **35**(5): 406-412.

Duggleby, R. G. (2006). "Domain relationships in thiamine diphosphate-dependent enzymes." Acc Chem Res **39**(8): 550-7.

Eddy, S. (1998). "HMMER." from http://hmmer.wustl.edu/.

Enders, D., O. Niemeier, et al. (2007). "Organocatalysis by N-heterocyclic, carbenes." Chemical Reviews **107**(12): 5606-5655.

Fang, R., P. F. Nixon, et al. (1998). "Identification of the catalytic glutamate in the E1 component of human pyruvate dehydrogenase." Febs Letters **437**(3): 273-277.

Fischer, M. and J. Pleiss (2003). "The Lipase Engineering Database: a navigation and analysis tool for protein families." Nucleic Acids Res **31**(1): 319-21.

Fischer, M., Q. K. Thai, et al. (2006). "DWARF--a data warehouse system for analyzing protein families." BMC Bioinformatics **7**: 495.

Frank, R. A., F. J. Leeper, et al. (2007). "Structure, mechanism and catalytic duality of thiamine-dependent enzymes." Cell Mol Life Sci **64**(7-8): 892-905.

Hawkins, C. F., A. Borges, et al. (1989). "A Common Structural Motif in Thiamin Pyrophosphate-Binding Enzymes." Febs Letters **255**(1): 77-82.

Iding, H., P. Siegert, et al. (1998). "Application of alpha-keto acid decarboxylases in biotransformations." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1385**(2): 307-322.

Jordan, F. (2003). "Current mechanistic understanding of thiamin diphosphatedependent enzymatic reactions." Natural Product Reports **20**(2): 184-201.

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-637.

Knoll, M., T. M. Hamm, et al. (2009). "The PHA Depolymerase Engineering Database: A systematic analysis tool for the diverse family of polyhydroxyalkanoate (PHA) depolymerases." BMC Bioinformatics **10**: 89.

Martin, J. I., W. C. Broaddus, et al. (2004). "A transcription factor decoy oligonucleotide that mimics the MMP-1 functional single nuclear polymorphism: A novel therapeutic for the inhibition of MMP-1 expression." Neuro-Oncology **6**(4): 333-333.

Mimori, K., H. Inoue, et al. (2002). "A single-nucleotide polymorphism of SMARCB1 in human breast cancers." Genomics **80**(3): 254-8.

Mueller, M., D. Gocke, et al. (2009). "Thiamin diphosphate in biological chemistry: exploitation of diverse thiamin diphosphate-dependent enzymes for asymmetric chemoenzymatic synthesis." FEBS Journal **276**(11).

PHYLODENDRON. "PHYLODENDRON." from http://iubio.bio.indiana.edu/treeapp/.

Podebrad, F., M. Heil, et al. (1999). "4,5-dimethyl-3-hydroxy-2[5H]-furanone (sotolone) - The odour of maple syrup urine disease." Journal of Inherited Metabolic Disease **22**(2): 107-114.

Pohl, M., G. A. Sprenger, et al. (2004). "A new perspective on thiamine catalysis." Curr Opin Biotechnol **15**(4): 335-42.

Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Russell, R. B. and G. J. Barton (1992). "Multiple Protein-Sequence Alignment from Tertiary Structure Comparison - Assignment of Global and Residue Confidence Levels." Proteins-Structure Function and Genetics **14**(2): 309-323.

Schellenberger, A. (1998). "Sixty years of thiamin diphosphate biochemistry." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1385**(2): 177-186.

Seifert, A. and J. Pleiss (2009). "Identification of selectivity-determining residues in cytochrome P450 monooxygenases: A systematic analysis of the substrate recognition site 5." Proteins-Structure Function and Bioinformatics **74**(4): 1028-1035.

Seifert, A., S. Vomund, et al. (2009). "Rational Design of a Minimal and Highly Enriched

CYP102A1 Mutant Library with Improved Regio-, Stereo- and Chemoselectivity." Chembiochem **10**(5): 853-861.

Sherry, S. T., M. H. Ward, et al. (2001). "dbSNP: the NCBI database of genetic variation." Nucleic Acids Research **29**(1): 308-311.

Shils, M. E. (2006). Modern Nutrition in Health and Disease (Modern Nutrition in Health & Disease, Lippincott Williams & Wilkins.

Stillger, T., M. Pohl, et al. (2006). "Reaction engineering of benzaldehyde lyase from Pseudomonas fluorescens catalyzing enantioselective C-C bond formation." Organic Process Research & Development **10**(6): 1172-1177.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Wang, J. J. L., P. R. Martin, et al. (1997). "Aspartate 155 of human transketolase is essential for thiamine diphosphate magnesium binding, and cofactor binding is required for dimer formation." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1341**(2): 165-172.

Wang, J. J. L., P. R. Martin, et al. (1997). "A transketolase assembly defect in a Wernicke-Korsakoff syndrome patient." Alcoholism-Clinical and Experimental Research **21**(4): 576-580.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics **17**(9): 847-8.

Zeitler, K. (2005). "Extending mechanistic routes in heterazolium catalysis-promising concepts for versatile synthetic methods." Angewandte Chemie-International Edition **44**(46): 7506-7510.

Zhao, J. and C. J. Zhong (2009). "A review on research progress of transketolase." Neurosci Bull **25**(2): 94-9

8.4.11. Figures

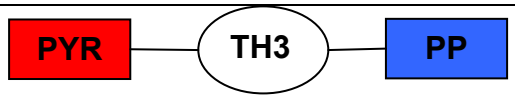
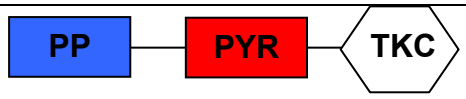
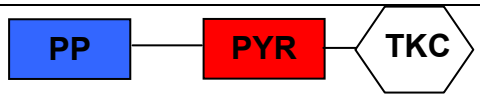
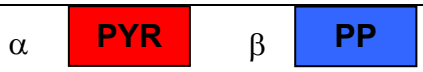
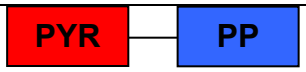
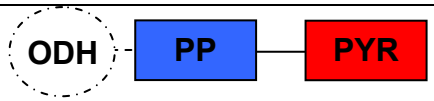
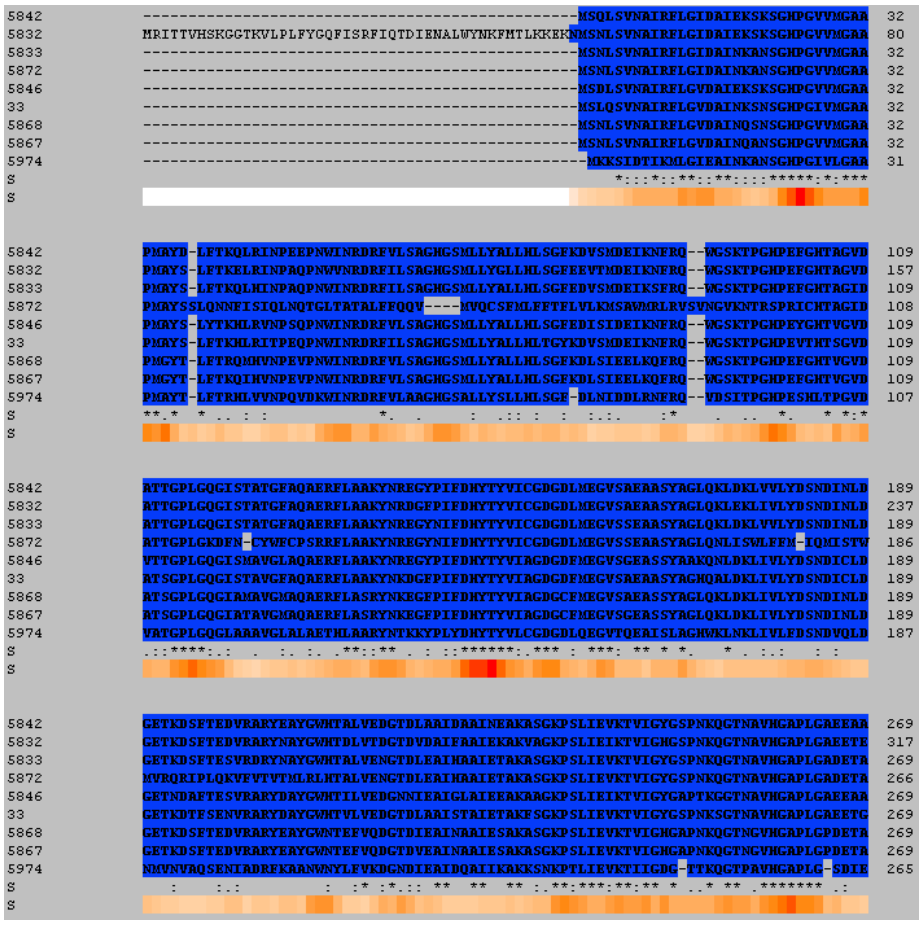
Superfamily ID	Superfamily	Structural arrangement of domains
1	DC	
2	TK	
3	OR	
4	K1	
5	K2	
6	SPDC	
7	PPDC	
8	KDH	

Figure 1 – Structural arrangement of protein domains of the superfamilies of the TEED.

All protein families are listed with their internal superfamily ID, the superfamily name and a 2D representation of the domain arrangement.



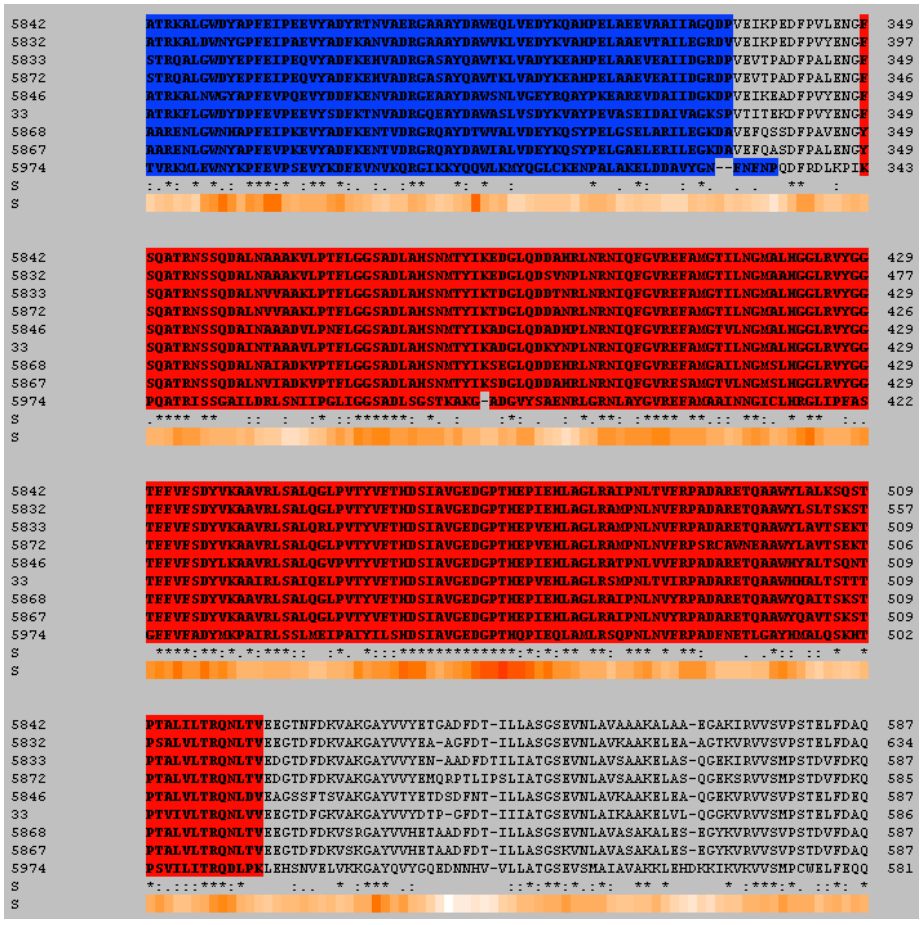


Figure 2 – Multisequence alignment of ThDP dependent proteins with annotated domains.

The pyrophosphate (PP) domain is coloured in blue, the pyrimidine (PYR) domain is coloured red. Annotated PP and PYR domains are available for all protein families in the TEED. The displayed multisequence alignment is taken from the transketolase homologous protein family (TEED ID 33)

8.4.12. Tables

Table 1. Sequences of human ThDP-dependent enzymes. Protein descriptions are taken from the protein GenBank (gi) entry. Isoforms of proteins are assigned to the same protein ID. Protein classifications state if the protein is considered a full sequence, a fragment of the full sequence or a SNP of the full sequence. Superfamily and homologous family describe the TEED identifiers, the homologous family id is given in brackets. Protein contains the name of the respective protein with the internal TEED protein id in brackets.

Superfamily	Homologous Family	Protein	gi	Sequence ID	Protein description from GenBank	Classification
DC	AHAS (11)	acetolactate synthase homolog (1433)	1730288	1886	acetolactate synthase homolog	Full sequence
DC	2-HPCL (21)	hydroxyacyl-CoA lyase (21)	20455027	21	isoform CRA_a	Full sequence
DC	2-HPCL (21)		119584656	3548	isoform CRA_b	Fragment
DC	2-HPCL (21)		193787013	1875	isoform CRA_c	Fragment
DC	2-HPCL (21)		6841208	1744	HSPC279	Fragment
DC	2-HPCL (21)		194378616	1777	unnamed protein product	Fragment
DC	2-HPCL (21)		194378068	1821	unnamed protein product	Fragment
DC	2-HPCL (21)		194376964	1770	unnamed protein product	Fragment
DC	2-HPCL (21)		6273457	1745	2-hydroxyphytanoyl-CoA lyase	Q447H,R543E
DC	2-HPCL (21)	unnamed (1317)	34531269	1741	unnamed protein product	Full sequence
DC	2-HPCL (21)	unnamed (1378)	194387780	1819	unnamed protein product	Full sequence
TK	TK (31)	Transketolase (31)	205277463	31	isoform 1	Full sequence
TK	TK (31)		193787540	5096	isoform 2	Fragment
TK	TK (31)		194381830	5080	unnamed protein product	Fragment
TK	TK (31)		38013966	5087	TKT protein	Fragment
TK	TK (31)		31417921	5108	TKT protein	Fragment
TK	TK (31)		14250367	5149	TKT protein	Fragment
TK	TK (31)		193787037	5193	unnamed protein product	Fragment
TK	TK (31)		62898960	5062	transketolase variant	Q367R
TK	TK (31)		388891	5065	transketolase	T585K, H586T, L587M
TK	TK (31)		37267	5066	transketolase	P426A
TK	TK (31)		194373693	5067	unnamed protein product	E374G

TK	TK (31)		194373793	5119	unnamed protein product	K145N,I378V
TK	TK (31)	Transketolase-like 2 (3748)	119625243	5124	Transketolase-like 2	Full sequence
TK	TK (31)		16553281	5235	unnamed protein product	Fragment
TK	TK (31)		189069449	5126	unnamed protein product	Y148H;I444T
TK	TK (31)		16552972	5127	unnamed protein product	F54L;M406I
TK	TK (31)		133777215	5128	Transketolase-like 2	I302V;P311H
TK	TK (31)		148744456	5120	Transketolase-like 2	Q590H
TK	TK (31)	Transketolase-like 1 (3781)	158257954	5162	isoform a	Full sequence
TK	TK (31)		221043878	5171	isoform b	Fragment
TK	TK (31)		55666480	5184	isoform c	Fragment
TK	TK (31)		221043730	5206	unnamed protein product	Y249C;H396Y
TK	TK (31)		158257880	5164	unnamed protein product	D26N
TK	TK (31)		34190015	5165	Transketolase-like 1	L24F;I152T
TK	TK (31)	Transketolase-like 1 (3791)	119593156	5178	transketolase-like 1, isoform CRA_b	Full sequence
TK	TK (31)	Transketolase-like 1 (3809)	119593155	5204	transketolase-like 1, isoform CRA_a	Full sequence
TK	TK (31)		119593159	5209	transketolase-like 1, isoform CRA_e	Fragment
TK	TK (31)	Transketolase-like 1 (3810)	119593157	5205	transketolase-like 1, isoform CRA_c	Full sequence
TK	TK (31)	Transketolase (3797)	1232175	5189	transketolase	Full sequence
TK	TK (31)	Transketolase-like 1 (3855)	122891454	5264	transketolase-like 1	Full sequence
K2	BCDH alpha (56)	2-oxoisovalerate dehydrogenase subunit alpha (56)	119577444	56	2-oxoisovalerate dehydrogenase subunit alpha	Full sequence
K2	BCDH alpha (56)		62089242	10805	branched chain keto acid dehydrogenase E1, alpha polypeptide variant	ΔR287
K2	BCDH alpha (56)		5705948	10841	branched-chain alpha-keto acid dehydrogenase complex E1 alpha subunit	V1G
K2	BCDH alpha (56)		189055345	10802	unnamed protein product	E377K
K2	BCDH alpha (56)		8176547	10803	branched-chain alpha-keto acid dehydrogenase E1 alpha subunit	V1G

K2	BCDH alpha (56)		386841	10806	branched-chain alpha-keto acid dehydrogenase	S35A
K2	BCDH alpha (56)		179360	10843	branched-chain alpha-keto acid dehydrogenase E1-alpha subunit	A181D
K2	BCDH alpha (56)	unnamed (8558)	194389886	10813	unnamed protein product	Full sequence
K2	BCDH alpha (56)	unnamed (8559)	34534581	10815	unnamed protein product	Full sequence
K2	BCDH alpha (56)	hypothetical (8572)	52545799	10850	hypothetical protein	Full sequence
K2						
K2	BCDH beta (57)	2-oxoisovalerate dehydrogenase subunit beta (57)	129034	57	2-oxoisovalerate dehydrogenase subunit beta	Full sequence
K2	BCDH beta (57)		221040270	11328	unnamed protein product	Fragment
K2	BCDH beta (57)		119569083	11439	branched chain keto acid dehydrogenase E1, beta polypeptide	Fragment
K2	BCDH beta (57)		747713	11309	unnamed protein product	T303S
K2	BCDH beta (57)		194385640	11329	unnamed protein product	Q276L
K2						
K2	AODH alpha (61)	mitochondrial PDHA1 (9896)	148357460	12435	mitochondrial PDHA1	Full sequence
K2	AODH alpha (61)		4505685	12453	pyruvate dehydrogenase (lipoamide) alpha 1 precursor	Fragment
K2	AODH alpha (61)		221041292	12530	unnamed protein product	Fragment
K2	AODH alpha (61)		62897039	12439	pyruvate dehydrogenase (lipoamide) alpha 1 variant	M282L
K2	AODH alpha (61)		62897537	12440	pyruvate dehydrogenase (lipoamide) alpha 1 variant	M282L;N328S
K2	AODH alpha (61)		189053388	12490	unnamed protein product	G278E
K2	AODH alpha (61)	pyruvate dehydrogenase E1-alpha precursor (9917)	387011	12463	pyruvate dehydrogenase E1-alpha precursor	Full sequence
K2	AODH alpha (61)	pyruvate dehydrogenase (9985)	119626468	12556	hCG1643458	Full sequence
K2	AODH alpha (61)		66267554	12549	PDHA2 protein	Fragment
	AODH alpha (61)		4885543	12546	pyruvate dehydrogenase (lipoamide) alpha 2	Fragment

8.4.13. Additional files

Table S1. Sequences of ThDP-dependent enzymes which were used to establish the TEED.

Sequences are annotated with accession numbers, organism of origin and references.

Family	Accession number (gi)	*Organism*	EC	Reference
		D C S u p e r f a m i l y		
POX	29337215	***Lactobacillus plantarum***	1.2.3.3	[1]
POX (Cyt.)	130693	***Escherichia coli***	1.2.2.2	[2]
IPDC	118333	***Enterobacter cloacae***	4.1.1.74	[3, 4]
PhePDC	6320588	***Saccharomyces cerevisiae***	4.1.1.-	[5]
PDC	515237	***Saccharomyces cerevisiae***	4.1.1.1	[6, 7]
	118391	***Zymomonas mobilis***	4.1.1.1	[8]
BFDC	3915757	***Pseudomonas putida***	4.1.1.7	[9, 10]
OCDC	730220	***Oxalobacter formigenes***	4.1.1.8	[11]
AHAS	124373	***Escherichia coli***	2.2.1.6	[12]
	33112641	***Escherichia coli***	2.2.1.6	[12]
	2507470	***Escherichia coli***	2.2.1.6	[12]
	124376	***Saccharomyces cerevisiae***	2.2.1.6	[13]
	124374	***Klebsiella pneumoniae***	2.2.1.6	[14]
	75172476	***Lolium multiflorum***	2.2.1.6	
BAL	1705519	***Pseudomonas fluorescens***	4.1.2.38	[15, 16]
CEAS	75488972	***Streptomyces clavuligerus***	-	[17]
GXC	84028422	***Escherichia coli***	4.1.1.47	[18]
CDP-ADS	unpublished	***Yersinia pseudotuberculosis***		[19, 20]
kdcA	75369656	***Lactococcus lactis***		[21]
SAAT	39932465	***Desulfonispora thiosulfatigenes***	2.3.3.15	[22]
2-HPCL	20455027	***Homo sapiens***		[23, 24]
	75174050	***Arabidopsis thaliana***		
SEPHCHC (MenD)	2507472	***Escherichia coli***	2.2.1.9	[25, 26]
	15790176	***Halobacterium sp.***	2.2.1.9	[26]
	12323219	***Arabidopsis thaliana***	2.2.1.9	[26]
CDH	185177534	***Azoarcus sp.***		[27]
THcHDOH	81687921	***Bacillus cereus***	3.7.1.n2	[28]
pigD	75361841	***Serratia marcescens***		[29]
		T K S u p e r f a m i l y		
TK	1351256	***Saccharomyces cerevisiae***	2.2.1.1	[30]
	1729976	***Homo sapiens***	2.2.1.1	[31]
	54042066	***Escherichia coli***	2.2.1.1	[32]

	169834000	***Streptococcus pneumoniae***	2.2.1.1	[33]
PK	15214330	***Bifidobacterium animalis***	4.1.2.9	[34]
	21363093	***Lactobacillus pentosus***	4.1.2.9	[35]
DHAS	108936021	***Pichia angusta***	2.2.1.3	[36]
DXPS	2501357	***Escherichia coli***	2.2.1.7	[37]
	81479889	***Fusobacterium nucleatum***	4.2.1.-	[38]
	122989637	***Candidatus Kuenenia stuttgartiensis***	2.2.1.7	[39]
O R S u p e r f a m i l y				
PFOR	75499539	***Desulfovibrio africanus***		[40, 41]
PFOR α	6685746	***Methanobacterium thermoautotrophicum***	1.2.7.1	[42]
PFOR β	6685734		1.2.7.1	[42]
PFOR γ	6685735		1.2.7.1	[42]
PFOR δ	6685747		1.2.7.1	[42]
KOR α	6685587	***Archaeoglobus fulgidus***	1.2.7.3	[43]
KOR β	6685588		1.2.7.3	[43]
KOR γ	6685586		1.2.7.3	[43]
KOR δ	74570221		1.2.7.3	[43]
VOR α	6686058	***Pyrococcus horikoshii***	1.2.7.7	[44]
VOR β	6686093		1.2.7.7	[44]
VOR γ	6685737		1.2.7.7	[44]
VOR δ	6686092		1.2.7.7	[44]
IOR α	62296914	***Pyrococcus kodakaraensis***	1.2.7.8	[45]
IOR β	62296919		1.2.7.8	[45]
K 1 S u p e r f a m i l y				
PDH	84027826	***Escherichia coli***	1.2.4.1	[46, 47]

K 2 S u p e r f a m i l y				
BCDH α	548403	***Homo sapiens***	1.2.4.4	[48]
BCDH β	129034		1.2.4.4	[48]
AODH α	113136	***Ralstonia eutropha***	1.1.1.-	[49]
AODH β	113137		1.1.1.-	[49]

S P D C S u p e r f a m i l y				
SPDC α	17432993	***Methanocaldococcus jannaschii***	4.1.1.79	[50]
SPDC β	17432994		4.1.1.79	[50]

P P D C S u p e r f a m i l y				
PPDC	22654224	***Streptomyces hygroscopicus***	4.1.1.82	[51]

K D H S u p e r f a m i l y				
OGDC	160395583	***Mycobacterium tuberculosis***	4.1.1.71	[52, 53]

DC

2-HPCL 2-hydroxyphytanoyl-CoA lyase, 2-hydroxyacyl-CoA lyase; **AHAS** acetohydroxyacid synthase; **BAL** benzaldehyde aldolase; **BFDC** benzoylformate decarboxylase; **CDH** cyclohexane- 1,2-dione hydrolase; **CDP-ADS (YerE)** CDP-4-aceto-3,6-dideoxygalactose synthase (YerE); **CEAS** N2-(2-carboxyethyl) arginine synthase; **GXC** glyoxylate carboligase; **IPDC** indolepyruvate decarboxylase; **OCDC** oxalyl-CoA decarboxylase; **PDC** pyruvate decarboxylase; **PhePDC** phenylpyruvate decarboxylase; **POX** pyruvate oxidase; **POX (Cyt)** pyruvate dehydrogenase [cytochrome]; **SAAT** sulfoacetaldehyde acetyltransferase; **SEPHCHC** 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid synthase; **THcHDOH** 3D-(3,5/4)-trihydroxycyclohexane-1,2-dione hydrolase; **pigD** protein pigD

TK

DHAS dihydroxyacetone synthase; **DXPS** 1-deoxy-D-xylulose-5-phosphate synthase; **PK** phosphoketolase (xylulose-5-phosphate/fructose-6-phosphate); **TK** transketolase

OR

IOR indolepyruvate ferredoxin oxidoreductase; **KOR** 2-keto(oxo)glutarate ferredoxin oxidoreductase; **PFOR** pyruvate ferredoxin oxidoreductase; **VOR** 2-keto(oxo)isovalerate ferredoxin oxidoreductase

K1

PDH pyruvate dehydrogenase E1 component

K2

BCDH branched-chain 2-ketoacid dehydrogenase; **AODH** acetoin dehydrogenase;

SPDC

SPDC sulfopyruvate decarboxylase

PPDC

PPDC phosphonopyruvate decarboxylase

KDH

OGDC 2-oxo(keto)glutarate decarboxylase, **OGDH** 2-oxoglutarate dehydrogenase

1. Muller YA, Schumacher G, Rudolph R, Schulz GE: **The refined structures of a stabilized mutant and of wild-type pyruvate oxidase from Lactobacillus plantarum**. *J Mol Biol* 1994, **237**(3):315-335.
2. Neumann P, Weidner A, Pech A, Stubbs MT, Tittmann K: **Structural basis for membrane binding and catalytic activation of the peripheral membrane enzyme pyruvate oxidase from Escherichia coli**. *Proc Natl Acad Sci U S A* 2008, **105**(45):17390-17395.
3. Schutz A, Sandalova T, Ricagno S, Hubner G, Konig S, Schneider G: **Crystal structure of thiamindiphosphate-dependent indolepyruvate decarboxylase from Enterobacter cloacae, an enzyme involved in the biosynthesis of the plant hormone indole-3-acetic acid**. *Eur J Biochem* 2003, **270**(10):2312-2321.
4. Schutz A, Golbik R, Tittmann K, Svergun DI, Koch MH, Hubner G, Konig S: **Studies on structure-function relationships of indolepyruvate decarboxylase from Enterobacter cloacae, a key enzyme of the indole acetic acid pathway**. *Eur J Biochem* 2003, **270**(10):2322-2331.
5. Vuralhan Z, Morais MA, Tai SL, Piper MD, Pronk JT: **Identification and characterization of phenylpyruvate decarboxylase genes in Saccharomyces cerevisiae**. *Appl Environ Microbiol* 2003, **69**(8):4534-4541.
6. Dyda F, Furey W, Swaminathan S, Sax M, Farrenkopf B, Jordan F: **Catalytic centers in the thiamin diphosphate dependent enzyme pyruvate decarboxylase at 2.4-A resolution**. *Biochemistry* 1993, **32**(24):6165-6170.
7. Rosche B, Breuer M, Hauer B, Rogers PL: **Screening of yeasts for cell-free production of (R)-phenylacetylcarbinol**. *Biotechnol Lett* 2003, **25**(11):841-845.
8. Dobritzsch D, Konig S, Schneider G, Lu G: **High resolution crystal structure of pyruvate decarboxylase from Zymomonas mobilis. Implications for substrate activation in pyruvate decarboxylases**. *J Biol Chem* 1998, **273**(32):20196-20204.
9. Hasson MS, Muscate A, McLeish MJ, Polovnikova LS, Gerlt JA, Kenyon GL, Petsko GA, Ringe D: **The crystal structure of benzoylformate decarboxylase at 1.6 A resolution: diversity of catalytic residues in thiamin diphosphate-dependent enzymes**. *Biochemistry* 1998, **37**(28):9918-9930.
10. Lingen B, Kolter-Jung D, Dunkelmann P, Feldmann R, Grotzinger J, Pohl M, Muller M: **Alteration of the substrate specificity of benzoylformate decarboxylase from Pseudomonas putida by directed evolution**. *Chembiochem* 2003, **4**(8):721-726.
11. Berthold CL, Moussatche P, Richards NG, Lindqvist Y: **Structural basis for activation of the thiamin diphosphate-dependent enzyme oxalyl-CoA decarboxylase by adenosine diphosphate**. *J Biol Chem* 2005, **280**(50):41645-41654.
12. Engel S, Vyazmensky M, Geresh S, Barak Z, Chipman DM: **Acetohydroxyacid synthase: a new enzyme for chiral synthesis of R-phenylacetylcarbinol**. *Biotechnol Bioeng* 2003, **83**(7):833-840.
13. Pang SS, Duggleby RG, Guddat LW: **Crystal structure of yeast acetohydroxyacid synthase: a target for herbicidal inhibitors**. *J Mol Biol* 2002, **317**(2):249-262.
14. Pang SS, Duggleby RG, Schowen RL, Guddat LW: **The crystal structures of Klebsiella pneumoniae acetolactate synthase with enzyme-bound cofactor and with an unusual intermediate**. *J Biol Chem* 2004, **279**(3):2242-2253.
15. Demir AS, Sesenoglu O, Dunkelmann P, Muller M: **Benzaldehyde lyase-catalyzed enantioselective carboligation of aromatic aldehydes with mono- and dimethoxy acetaldehyde**. *Org Lett* 2003, **5**(12):2047-2050.
16. Maraite A, Schmidt T, Ansorge-Schumacher MB, Brzozowski AM, Grogan G: **Structure of the ThDP-dependent enzyme benzaldehyde lyase refined to 1.65 A**

resolution. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2007, **63**(Pt 7):546-548.

17. Caines ME, Elkins JM, Hewitson KS, Schofield CJ: **Crystal structure and mechanistic implications of N2-(2-carboxyethyl)arginine synthase, the first enzyme in the clavulanic acid biosynthesis pathway**. *J Biol Chem* 2004, **279**(7):5685-5692.
18. Kaplun A, Binshtein E, Vyazmensky M, Steinmetz A, Barak Z, Chipman DM, Tittmann K, Shaanan B: **Glyoxylate carboligase lacks the canonical active site glutamate of thiamine-dependent enzymes**. *Nature Chemical Biology* 2008, **4**(2):113-118.
19. Chen HW, Guo ZH, Liu HW: **Biosynthesis of yersiniose: Attachment of the two-carbon branched-chain is catalyzed by a thiamine pyrophosphate-dependent flavoprotein**. *Journal of the American Chemical Society* 1998, **120**(45):11796-11797.
20. Mansoorabadi SO, Thibodeaux CJ, Liu HW: **The diverse roles of flavin coenzymes--nature's most versatile thespians**. *J Org Chem* 2007, **72**(17):6329-6342.
21. Berthold CL, Gocke D, Wood D, Leeper FJ, Pohl M, Schneider G: **Structure of the branched-chain keto acid decarboxylase (KdcA) from Lactococcus lactis provides insights into the structural basis for the chemoselective and enantioselective carboligation reaction**. *Acta Crystallographica Section D-Biological Crystallography* 2007, **63**:1217-1224.
22. Ruff J, Denger K, Cook AM: **Sulphoacetaldehyde acetyltransferase yields acetyl phosphate: purification from Alcaligenes defragrans and gene clusters in taurine degradation**. *Biochem J* 2003, **369**(Pt 2):275-285.
23. Foulon V, Antonenkov VD, Croes K, Waelkens E, Mannaerts GP, Van Veldhoven PP, Casteels M: **Purification, molecular cloning, and expression of 2-hydroxyphytanoyl-CoA lyase, a peroxisomal thiamine pyrophosphate-dependent enzyme that catalyzes the carbon-carbon bond cleavage during alpha-oxidation of 3-methyl-branched fatty acids**. *Proc Natl Acad Sci U S A* 1999, **96**(18):10039-10044.
24. Casteels M, Foulon V, Mannaerts GP, Van Veldhoven PP: **Alpha-oxidation of 3-methyl-substituted fatty acids and its thiamine dependence**. *European Journal of Biochemistry* 2003, **270**(8):1619-1627.
25. Dawson A, Fyfe PK, Hunter WN: **Specificity and reactivity in menaquinone biosynthesis: the structure of Escherichia coli MenD (2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexadiene-1-carboxylate synthase)**. *J Mol Biol* 2008, **384**(5):1353-1368.
26. Bhasin M, Billinsky JL, Palmer DR: **Steady-state kinetics and molecular evolution of Escherichia coli MenD [(1R,6R)-2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase], an anomalous thiamin diphosphate-dependent decarboxylase-carboligase**. *Biochemistry* 2003, **42**(46):13496-13504.
27. Harder J: **Anaerobic degradation of cyclohexane-1,2-diol by a new Azoarcus species**. *Arch Microbiol* 1997, **168**(3):199-204.
28. Yoshida K, Yamaguchi M, Morinaga T, Kinehara M, Ikeuchi M, Ashida H, Fujita Y: **myo-Inositol catabolism in Bacillus subtilis**. *J Biol Chem* 2008, **283**(16):10415-10424.
29. Williamson NR, Simonsen HT, Ahmed RA, Goldet G, Slater H, Woodley L, Leeper FJ, Salmond GP: **Biosynthesis of the red antibiotic, prodigiosin, in Serratia: identification of a novel 2-methyl-3-n-amyl-pyrrole (MAP) assembly pathway, definition of the terminal condensing enzyme, and implications for undecylprodigiosin biosynthesis in Streptomyces**. *Mol Microbiol* 2005, **56**(4):971-989.

30. Nikkola M, Lindqvist Y, Schneider G: **Refined structure of transketolase from Saccharomyces cerevisiae at 2.0 A resolution**. *J Mol Biol* 1994, **238**(3):387-404.
31. Obiol-Pardo C, Rubio-Martinez J: **Homology modeling of human transketolase: description of critical sites useful for drug design and study of the cofactor binding mode**. *J Mol Graph Model* 2009, **27**(6):723-734.
32. Asztalos P, Parthier C, Golbik R, Kleinschmidt M, Hubner G, Weiss MS, Friedemann R, Wille G, Tittmann K: **Strain and near attack conformers in enzymic thiamin catalysis: X-ray crystallographic snapshots of bacterial transketolase in covalent complex with donor ketoses xylulose 5-phosphate and fructose 6-phosphate, and in noncovalent complex with acceptor aldose ribose 5-phosphate**. *Biochemistry* 2007, **46**(43):12037-12052.
33. Reizer J, Reizer A, Bairoch A, Saier MH, Jr.: **A diverse transketolase family that includes the RecP protein of Streptococcus pneumoniae, a protein implicated in genetic recombination**. *Res Microbiol* 1993, **144**(5):341-347.
34. Meile L, Rohr LM, Geissman TA, Herensperger M, Teuber M: **Characterization of the D-xylulose 5-phosphate/D-Fructose 6-phosphate phosphoketolase gene (xfp) from Bifidobacterium lactis**. *Journal of Bacteriology* 2001, **183**(9):2929-2936.
35. Posthuma CC, Bader R, Engelmann R, Postma PW, Hengstenberg W, Pouwels PH: **Expression of the xylulose 5-phosphate phosphoketolase gene, xpkA, from Lactobacillus pentosus MD363 is induced by sugars that are fermented via the phosphoketolase pathway and is repressed by glucose mediated by CcpA and the mannose phosphoenolpyruvate phosphotransferase system**. *Applied and Environmental Microbiology* 2002, **68**(2):831-837.
36. Janowicz ZA, Eckart MR, Drewke C, Roggenkamp RO, Hollenberg CP, Maat J, Ledeboer AM, Visser C, Verrips CT: **Cloning and characterization of the DAS gene encoding the major methanol assimilatory enzyme from the methylotrophic yeast hansenula-polymorpha**. *Nucleic Acids Research* 1985, **13**(9):3043-3062.
37. Xiang S, Usunow G, Lange G, Busch M, Tong L: **Crystal structure of 1-deoxy-d-xylulose 5-phosphate synthase, a crucial enzyme for isoprenoids biosynthesis**. *Journal of Biological Chemistry* 2007, **282**(4):2676-2682.
38. Kapatral V, Anderson I, Ivanova N, Reznik G, Los T, Lykidis A, Bhattacharyya A, Bartman A, Gardner W, Grechkin G *et al*: **Genome sequence and analysis of the oral bacterium Fusobacterium nucleatum strain ATCC 25586**. *J Bacteriol* 2002, **184**(7):2005-2018.
39. Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, Horn M, Daims H, Bartol-Mavel D, Wincker P *et al*: **Deciphering the evolution and metabolism of an anammox bacterium from a community genome**. *Nature* 2006, **440**(7085):790-794.
40. Chabriere E, Vernede C, Guigliarelli B, Charon MH, Hatchikian EC, Fontecilla-Camps JC: **Crystal structure of the free radical intermediate of pyruvate : ferredoxin oxidoreductase**. *Science* 2001, **294**(5551):2559-2563.
41. Chabriere E, Charon MH, Volbeda A, Pieulle L, Hatchikian EC, Fontecilla-Camps JC: **Crystal structures of the key anaerobic enzyme pyruvate : ferredoxin oxidoreductase, free and in complex with pyruvate**. *Nature Structural Biology* 1999, **6**(2):182-190.
42. Smith DR, DoucetteStamm LA, Deloughery C, Lee HM, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K *et al*: **Complete genome sequence of Methanobacterium thermoautotrophicum Delta H: Functional analysis and comparative genomics**. *Journal of Bacteriology* 1997, **179**(22):7135-7155.
43. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD *et al*: **The complete genome sequence of the**

hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature* 1997, **390**(6658):364-&.

44. Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A *et al*: **Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3**. *DNA Res* 1998, **5**(2):55-76.
45. Siddiqui MA, Fujiwara S, Imanaka T: **Indolepyruvate ferredoxin oxidoreductase from Pyrococcus sp. KOD1 possesses a mosaic structure showing features of various oxidoreductases**. *Mol Gen Genet* 1997, **254**(4):433-439.
46. Stephens PE, Darlison MG, Lewis HM, Guest JR: **The pyruvate-dehydrogenase complex of Escherichia-coli-K12 - nucleotide-sequence encoding the pyruvate-dehydrogenase component**. *European Journal of Biochemistry* 1983, **133**(1):155-162.
47. Arjunan P, Chandrasekhar K, Sax M, Brunskill A, Nemeria N, Jordan F, Furey W: **Structural determinants of enzyme binding affinity: The E1 component of pyruvate dehydrogenase from Escherichia coli in complex with the inhibitor thiamin thiazolone diphosphate**. *Biochemistry* 2004, **43**(9):2405-2411.
48. Wynn RM, Kato M, Machius M, Chuang JL, Li J, Tomchick DR, Chuang DT: **Molecular mechanism for regulation of the human mitochondrial branched-chain alpha-ketoacid dehydrogenase complex by phosphorylation**. *Structure* 2004, **12**(12):2185-2196.
49. Priefert H, Hein S, Kruger N, Zeh K, Schmidt B, Steinbuchel A: **Identification and molecular characterization of the Alcaligenes eutrophus H16 aco operon genes involved in acetoin catabolism**. *J Bacteriol* 1991, **173**(13):4056-4071.
50. Graupner M, Xu H, White RH: **Identification of the gene encoding sulfopyruvate decarboxylase, an enzyme involved in biosynthesis of coenzyme M**. *J Bacteriol* 2000, **182**(17):4862-4867.
51. Nakashita H, Kozuka K, Hidaka T, Hara O, Seto H: **Identification and expression of the gene encoding phosphonopyruvate decarboxylase of Streptomyces hygroscopicus**. *Biochim Biophys Acta* 2000, **1490**(1-2):159-162.
52. Tian J, Bryk R, Itoh M, Suematsu M, Nathan C: **Variant tricarboxylic acid cycle in Mycobacterium tuberculosis: identification of alpha-ketoglutarate decarboxylase**. *Proc Natl Acad Sci U S A* 2005, **102**(30):10670-10675.
53. Tian J, Bryk R, Shi S, Erdjument-Bromage H, Tempst P, Nathan C: **Mycobacterium tuberculosis appears to lack alpha-ketoglutarate dehydrogenase and encodes pyruvate dehydrogenase in widely separated genes**. *Mol Microbiol* 2005, **57**(3):859-868.

9. Gesamtliteraturverzeichnis

Ahamed, T., S. Chilamkurthi, B. K. Nfor, P. D. E. M. Verhaert, G. W. K. van Dedem, et al. (2008). "Selection of pH-related parameters in ion-exchange chromatography using pH-gradient operations." Journal of Chromatography A **1194**(1): 22-29.

Ahamed, T., B. K. Nfor, P. D. Verhaert, G. W. van Dedem, L. A. van der Wielen, et al. (2007). "pH-gradient ion-exchange chromatography: an analytical tool for design and optimization of protein separations." J Chromatogr A **1164**(1-2): 181-8.

Ahamed, T., M. Ottens, B. K. Nfor, G. W. K. van Dedem and L. A. M. van der Wielen (2006). "A generalized approach to thermodynamic properties of biomolecules for use in bioseparation process design." Fluid Phase Equilibria **241**(1-2): 268-282.

Akoh, C. C., G. C. Lee and J. F. Shaw (2004). "Protein engineering and applications of Candida rugosa lipase isoforms." Lipids **39**(6): 513-526.

Alberts, B., A. Johnson, P. Walter, J. Lewis, M. Raff, et al. (2008). Molecular Biology of the Cell, Taylor & Francis.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Bairoch, A., L. Bougueleret, S. Altairac, V. Amendolia, A. Auchincloss, et al. (2008). "The Universal Protein Resource (UniProt)." Nucleic Acids Research **36**: D190-D195.

Barth, S., M. Fischer, R. D. Schmid and J. Pleiss (2004). "The database of epoxide hydrolases and haloalkane dehalogenases: one structure, many functions." Bioinformatics **20**(16): 2845-2847.

Barth, S., M. Fischer, R. D. Schmid and J. Pleiss (2004). "Sequence and structure of epoxide hydrolases: a systematic analysis." Proteins **55**(4): 846-55.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2009). "GenBank." Nucleic Acids Research **37**: D26-D31.

Berman, H. M., T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, et al. (2002). "The Protein Data Bank." Acta Crystallogr D Biol Crystallogr **58**(Pt 6 No 1): 899-907.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, et al. (2000). "The Protein Data Bank." Nucleic Acids Research **28**(1): 235-242.

Berry, M. J., R. M. Tujebajeva, P. R. Copeland, X. M. Xu, B. A. Carlson, et al. (2001). "Selenocysteine incorporation directed from the 3'UTR: characterization of eukaryotic EFsec and mechanistic implications." Biofactors **14**(1-4): 17-24.

Berthold, C. L., D. Gocke, D. Wood, F. J. Leeper, M. Pohl, et al. (2007). "Structure of the branched-chain keto acid decarboxylase (KdcA) from Lactococcus lactis provides insights into the structural basis for the chemoselective and enantioselective carboligation reaction." Acta Crystallographica Section D-Biological Crystallography **63**: 1217-1224.

Bornemann, S., D. H. G. Crout, H. Dalton, D. W. Hutchinson, G. Dean, et al. (1993). "Stereochemistry of the Formation of Lactaldehyde and Acetoin Produced by the Pyruvate Decarboxylases of Yeast (Saccharomyces Sp) and Zymomonas-Mobilis - Different Boltzmann Distributions between Bound Forms of the Electrophile, Acetaldehyde, in the 2 Enzymatic-Reactions." Journal of the Chemical Society-Perkin Transactions 1(3): 309-311.

Brunak, S. and J. Engelbrecht (1996). "Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level." Proteins **25**(2): 237-52.

Candy, J. M. and R. G. Duggleby (1998). "Structure and properties of pyruvate decarboxylase and site-directed mutagenesis of the Zymomonas mobilis enzyme." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1385**(2): 323-338.

Casteels, M., V. Foulon, G. P. Mannaerts and P. P. Van Veldhoven (2003). "Alpha-oxidation of 3-methyl-substituted fatty acids and its thiamine dependence." European Journal of Biochemistry **270**(8): 1619-1627.

Chou, T. and G. Lakatos (2004). "Clustered bottlenecks in mRNA translation and protein synthesis." Phys Rev Lett **93**(19): 198101.

Clarke, J., E. Cota, S. B. Fowler and S. J. Hamill (1999). "Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway." Structure **7**(9): 1145-53.

Codd, E. F. (1970). "A relational model of data for large shared data banks. 1970." MD Comput **15**(3): 162-6.

Cortazzo, P., C. Cervenansky, M. Marin, C. Reiss, R. Ehrlich, et al. (2002). "Silent mutations affect in vivo protein folding in Escherichia coli." Biochem Biophys Res Commun **293**(1): 537-41.

Costelloe, S. J., J. M. Ward and P. A. Dalby (2008). "Evolutionary analysis of the TPP-dependent enzyme family." J Mol Evol **66**(1): 36-49.

Crick, F. H. (1966). "Codon--anticodon pairing: the wobble hypothesis." J Mol Biol **19**(2): 548-55.

Cromartie, T. H. and C. T. Walsh (1976). "Escherichia-Coli Glyoxalate Carboligase - Properties and Reconstitution with 5-Deazafad and 1,5-Dihydrodeazafadh2." Journal of Biological Chemistry **251**(2): 329-333.

Crombie, T., J. C. Swaffield and A. J. Brown (1992). "Protein folding within the cell is influenced by controlled rates of polypeptide elongation." J Mol Biol **228**(1): 7-12.

Curran, J. F. and M. Yarus (1989). "Rates of aminoacyl-tRNA selection at 29 sense codons in vivo." J Mol Biol **209**(1): 65-77.

de Maria, P. D., C. Carboni-Oerlemans, B. Tuin, G. Bargeman, A. van der Meer, et al. (2005). "Biotechnological applications of Candida antarctica lipase A: State-of-the-art." Journal of Molecular Catalysis B-Enzymatic **37**(1-6): 36-46.

Demir, A. S., P. Ayhan and S. B. Sopaci (2007). "Thiamine pyrophosphate dependent enzyme catalyzed reactions: Stereoselective C-Cbond formations in water." Clean-Soil Air Water **35**(5): 406-412.

Dolinsky, T. J., P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, et al. (2007). "PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations." Nucleic Acids Res **35**(Web Server issue): W522-5.

dos Reis, M., R. Savva and L. Wernisch (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic Acids Res **32**(17): 5036-44.

Duggleby, R. G. (2006). "Domain relationships in thiamine diphosphate-dependent enzymes." Acc Chem Res **39**(8): 550-7.

Eddy, S. (1998). "HMMER." from http://hmmer.wustl.edu/.

Enders, D., O. Niemeier and A. Henseler (2007). "Organocatalysis by N-heterocyclic, carbenes." Chemical Reviews **107**(12): 5606-5655.

Ericsson, D. J., A. Kasrayan, P. Johanssonl, T. Bergfors, A. G. Sandstrom, et al. (2008). "X-ray structure of Candida antarctica lipase a shows A novel lid structure and a likely mode of interfacial activation." Journal of Molecular Biology **376**(1): 109-119.

Fang, R., P. F. Nixon and R. G. Duggleby (1998). "Identification of the catalytic glutamate in the E1 component of human pyruvate dehydrogenase." Febs Letters **437**(3): 273-277.

Fenske, C., G. J. Palm and W. Hinrichs (2003). "How unique is the genetic code?" Angewandte Chemie-International Edition **42**(6): 606-610.

Firebird. "Firebird." from http://sourceforge.net/projects/firebird.

Fischer, M., M. Knoll, D. Sirim, F. Wagner, S. Funke, et al. (2007). "The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family." Bioinformatics **23**(15): 2015-2017.

Fischer, M. and J. Pleiss (2003). "The Lipase Engineering Database: a navigation and analysis tool for protein families." Nucleic Acids Res **31**(1): 319-21.

Fischer, M., Q. K. Thai, M. Grieb and J. Pleiss (2006). "DWARF--a data warehouse system for analyzing protein families." BMC Bioinformatics **7**: 495.

Frank, R. A., F. J. Leeper and B. F. Luisi (2007). "Structure, mechanism and catalytic duality of thiamine-dependent enzymes." Cell Mol Life Sci **64**(7-8): 892-905.

Gaasterland, T. (1998). "Structural genomics taking shape." Trends in Genetics **14**(4): 135-135.

Garel, J. P., G. Chavancy, A. Chevallier, A. Fournier, G. Marbaix, et al. (1981). "[tRNA adaptation and the optimization of translation]." Reprod Nutr Dev **21**(2): 177-83.

Gonnet, G. H., M. A. Cohen and S. A. Benner (1992). "Exhaustive matching of the entire protein sequence database." Science **256**(5062): 1443-5.

Gotor-Fernandez, V., E. Busto and V. Gotor (2006). "Candida antarctica lipase B: An ideal biocatalyst for the preparation of nitrogenated organic compounds." Advanced Synthesis & Catalysis **348**(7-8): 797-812.

Gotte, G., M. Libonati and D. V. Laurents (2003). "Glycosylation and specific deamidation of ribonuclease B affect the formation of three-dimensional domain-swapped oligomers." J Biol Chem **278**(47): 46241-51.

Grantham, R., C. Gautier and M. Gouy (1980). "Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type." Nucleic Acids Res **8**(9): 1893-912.

Grimsley, G. R., J. M. Scholtz and C. N. Pace (2009). "A summary of the measured pK values of the ionizable groups in folded proteins." Protein Science **18**(1): 247-251.

Grochulski, P., Y. Li, J. D. Schrag, F. Bouthillier, P. Smith, et al. (1993). "Insights into interfacial activation from an open structure of Candida rugosa lipase." J Biol Chem **268**(17): 12843-7.

Grochulski, P., Y. Li, J. D. Schrag and M. Cygler (1994). "Two conformational states of Candida rugosa lipase." Protein Sci **3**(1): 82-91.

Hale, R. S. and G. Thompson (1998). "Codon optimization of the gene encoding a domain from human type 1 neurofibromin protein results in a threefold improvement in expression level in Escherichia coli." Protein Expr Purif **12**(2): 185-8.

Hallgren, E., F. Kalman, D. Farnan, C. Horvath and J. Stahlberg (2000). "Protein retention in ion-exchange chromatography: effect of net charge and charge distribution." J Chromatogr A **877**(1-2): 13-24.

Hannig, G. and S. C. Makrides (1998). "Strategies for optimizing heterologous protein expression in Escherichia coli." Trends Biotechnol **16**(2): 54-60.

Hawkins, C. F., A. Borges and R. N. Perham (1989). "A Common Structural Motif in Thiamin Pyrophosphate-Binding Enzymes." Febs Letters **255**(1): 77-82.

Healthcare, G. (2004). Ion Exchange Chromatography & Chromatofocusing:Principles and Methods, GE Healthcare.

Hoekema, A., R. A. Kastelein, M. Vasser and H. A. de Boer (1987). "Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression." Mol Cell Biol **7**(8): 2914-24.

Holm, L. and C. Sander (1995). "Dali: a network tool for protein structure comparison." Trends Biochem Sci **20**(11): 478-80.

Holmquist, M. (2000). "Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms." Curr Protein Pept Sci **1**(2): 209-35.

Hua, Z., H. Wang, D. Chen, Y. Chen and D. Zhu (1994). "Enhancement of expression of human granulocyte-macrophage colony stimulating factor by argU gene product in Escherichia coli." Biochem Mol Biol Int **32**(3): 537-43.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, et al. (2009). "InterPro: the integrative protein signature database." Nucleic Acids Research **37**: D211-D215.

Iding, H., P. Siegert, K. Mesch and M. Pohl (1998). "Application of alpha-keto acid decarboxylases in biotransformations." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1385**(2): 307-322.

Ikemura, T. (1981). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system." J Mol Biol **151**(3): 389-409.

Ikemura, T. (1985). "Codon usage and tRNA content in unicellular and multicellular organisms." Mol Biol Evol **2**(1): 13-34.

Inmon, W. H. (2002). Building the Data Warehouse, Wiley & Sons.

Jamison, D. C. (2003). "Structured Query Language (SQL) fundamentals." Curr Protoc Bioinformatics **Chapter 9**: Unit9 2.

Jordan, F. (2003). "Current mechanistic understanding of thiamin diphosphatedependent enzymatic reactions." Natural Product Reports **20**(2): 184-201.

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-637.

Kersey, P., L. Bower, L. Morris, A. Horne, R. Petryszak, et al. (2005). "Integr8 and Genome Reviews: integrated views of complete genomes and proteomes." Nucleic Acids Res **33**(Database issue): D297-302.

Kim, E. Y., K. H. Oh, M. H. Lee, C. H. Kang, T. K. Oh, et al. (2009). "Novel cold-adapted alkaline lipase from an intertidal flat metagenome and proposal for a new family of bacterial lipases." Appl Environ Microbiol **75**(1): 257-60.

Kirk, O. and M. W. Christensen (2002). "Lipases from Candida antarctica: Unique Biocatalysts from a Unique Origin." Org. Proc. Res. **6**(4): 446–451.

Knoll, M., T. M. Hamm, F. Wagner, V. Martinez and J. Pleiss (2009). "The PHA Depolymerase Engineering Database: A systematic analysis tool for the diverse family of polyhydroxyalkanoate (PHA) depolymerases." BMC Bioinformatics **10**: 89.

Knoll, M. and J. Pleiss (2008). "The Medium-Chain Dehydrogenase/Reductase Engineering Database: A systematic analysis of a diverse protein family to understand sequence-structure-function relationship." Protein Sci.

Komar, A. A., T. Lesnik and C. Reiss (1999). "Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation." FEBS Lett **462**(3): 387-91.

Kopelman, M. D., A. D. Thomson, I. Guerrini and E. J. Marshall (2009). "The Korsakoff syndrome: clinical aspects, psychology and treatment." Alcohol Alcohol **44**(2): 148-54.

Koschorreck, M., M. Fischer, S. Barth and J. Pleiss (2005). "How to find soluble proteins: a comprehensive analysis of alpha/beta hydrolases for recombinant expression in E. coli." BMC Genomics **6**(1): 49.

Kragelund, B. B., P. Hojrup, M. S. Jensen, C. K. Schjerling, E. Juul, et al. (1996). "Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family." J Mol Biol **256**(1): 187-200.

Lammle, K., H. Zipper, M. Breuer, B. Hauer, C. Buta, et al. (2007). "Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning." J Biotechnol **127**(4): 575-92.

Li, X. Q., L. F. Luo and C. Q. Liu (2003). "[The relation between translation speed and protein secondary structure]." Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai) **35**(2): 193-6.

Lindqvist, Y., G. Schneider, U. Ermler and M. Sundstrom (1992). "3-Dimensional Structure of Transketolase, a Thiamine Diphosphate Dependent Enzyme, at 2.5 Angstrom Resolution." Embo Journal **11**(7): 2373-2379.

Makhoul, C. H. and E. N. Trifonov (2002). "Distribution of rare triplets along mRNA and their relation to protein folding." J Biomol Struct Dyn **20**(3): 413-20.

Makoff, A. J., M. D. Oxer, M. A. Romanos, N. F. Fairweather and S. Ballantine (1989). "Expression of tetanus toxin fragment C in E. coli: high level expression by removing rare codons." Nucleic Acids Res **17**(24): 10191-202.

Malmquist, G., U. H. Nilsson, M. Norrman, U. Skarp, M. Stromgren, et al. (2006). "Electrostatic calculations and quantitative protein retention models for ion exchange chromatography." J Chromatogr A **1115**(1-2): 164-86.

Martin, J. I., W. C. Broaddus and H. I. Fillmore (2004). "A transcription factor decoy oligonucleotide that mimics the MMP-1 functional single nuclear polymorphism: A novel therapeutic for the inhibition of MMP-1 expression." Neuro-Oncology **6**(4): 333-333.

Martinelle, M., M. Holmquist and K. Hult (1995). "On the interfacial activation of Candida antarctica lipase A and B as compared with Humicola lanuginosa lipase." Biochim Biophys Acta **1258**(3): 272-6.

Mattes, R. (1993). Principles of Gene Expression. Biotechnology, A Multi-Volume comprehensive Treatise. **2**.

Mattes, R. (2001). "The production of improved tissue-type plasminogen activator in Escherichia coli." Semin Thromb Hemost **27**(4): 325-36.

Melander, W. R., Z. el Rassi and C. Horvath (1989). "Interplay of hydrophobic and electrostatic interactions in biopolymer chromatography. Effect of salts on the retention of proteins." J Chromatogr **469**: 3-27.

Mimori, K., H. Inoue, T. Shiraishi, H. Ueo, K. Mafune, et al. (2002). "A single-nucleotide polymorphism of SMARCB1 in human breast cancers." Genomics **80**(3): 254-8.

Mitra, C. K. and M. Rani (1993). "Protein Sequences as Random Fractals." Journal of Biosciences **18**(2): 213-220.

Mosbacher, T. G., M. Mueller and G. E. Schulz (2005). "Structure and mechanism of the ThDP-dependent benzaldehyde lyase from Pseudomonas fluorescens." Febs J **272**(23): 6067-76.

Mueller, M., D. Gocke and M. Pohl (2009). "Thiamin diphosphate in biological chemistry: exploitation of diverse thiamin diphosphate-dependent enzymes for asymmetric chemoenzymatic synthesis." FEBS Journal **276**(11).

Musto, H., H. Romero and A. Zavala (2003). "Translational selection is operative for synonymous codon usage in Clostridium perfringens and Clostridium acetobutylicum." Microbiology **149**(Pt 4): 855-63.

Nakamura, Y., T. Gojobori and T. Ikemura (2000). "Codon usage tabulated from international DNA sequence databases: status for the year 2000." Nucleic Acids Res **28**(1): 292.

Noh, H., S. T. Yohe and E. A. Vogler (2008). "Volumetric interpretation of protein adsorption: Ion-exchange adsorbent capacity, protein pI, and interaction energetics." Biomaterials **29**(13): 2033-48.

Ollis, D. L., E. Cheah, M. Cygler, B. Dijkstra, F. Frolow, et al. (1992). "The alpha/beta hydrolase fold." Protein Eng **5**(3): 197-211.

Orengo, C. A., D. T. Jones and J. M. Thornton (1994). "Protein superfamilies and domain superfolds." Nature **372**(6507): 631-4.

Orrenius, C., T. Norin, K. Hult and G. Carrea (1995). "The Candida antarctica lipase B catalysed kinetic resolution of seudenol in non-aqueous media of controlled water activity." Tetrahedron-Asymmetry **6**(12): 3023-3030.

Orrenius, C., N. Ohrner, D. Rotticci, A. Mattson, K. Hult, et al. (1995). "Candida-Antarctica Lipase-B Catalyzed Kinetic Resolutions - Substrate Structure Requirements for the Preparation of Enantiomerically Enriched Secondary Alcanols." Tetrahedron-Asymmetry **6**(5): 1217-1220.

Pabst, T. M., G. Carta, N. Ramasubramanyan, A. K. Hunter, P. Mensah, et al. (2008). "Separation of Protein Charge Variants with Induced pH Gradients Using Anion Exchange Chromatographic Columns." Biotechnology Progress **24**(5): 1096-1106.

Palekar, A. A., P. T. Vasudevan and S. Yan (2000). "Purification of lipase: A review." Biocatalysis and Biotransformation **18**(3): 177-200.

Patrickios, C. S. and E. N. Yamasaki (1995). "Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory." Anal Biochem **231**(1): 82-91.

Pedersen, S. (1984). "Escherichia coli ribosomes translate in vivo with variable rate." Embo J **3**(12): 2895-8.

PERL. "PERL." from http://www.perl.org/.

Pfeffer, J., S. Richter, J. Nieveler, C. E. Hansen, R. B. Rhlid, et al. (2006). "High yield expression of Lipase A from Candida antarctica in the methylotrophic yeast Pichia pastoris and its purification and characterisation." Applied Microbiology and Biotechnology **72**(5): 931-938.

PHYLODENDRON. "PHYLODENDRON." from http://iubio.bio.indiana.edu/treeapp/.

Pleiss, J., M. Fischer, M. Peiker, C. Thiele and R. D. Schmid (2000). "Lipase engineering database - Understanding and exploiting sequence-structure-function relationships." Journal of Molecular Catalysis B-Enzymatic **10**(5): 491-508.

Podebrad, F., M. Heil, S. Reichert, A. Mosandl, A. C. Sewell, et al. (1999). "4,5-dimethyl-3-hydroxy-2[5H]-furanone (sotolone) - The odour of maple syrup urine disease." Journal of Inherited Metabolic Disease **22**(2): 107-114.

Pohl, M., G. A. Sprenger and M. Muller (2004). "A new perspective on thiamine catalysis." Curr Opin Biotechnol **15**(4): 335-42.

Purvis, I. J., A. J. Bettany, T. C. Santiago, J. R. Coggins, K. Duncan, et al. (1987). "The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis." J Mol Biol **193**(2): 413-7.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, et al. (2005). "InterProScan: protein domains identifier." Nucleic Acids Research **33**: W116-W120.

Rangwala, S. H., R. F. Finn, C. E. Smith, S. A. Berberich, W. J. Salsgiver, et al. (1992). "High-level production of active HIV-1 protease in Escherichia coli." Gene **122**(2): 263-9.

Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Richardson, J. S. (1981). "The anatomy and taxonomy of protein structure." Adv Protein Chem **34**: 167-339.

Ropson, I. J., B. C. Yowler, P. M. Dalessio, L. Banaszak and J. Thompson (2000). "Properties and crystal structure of a beta-barrel folding mutant." Biophys J **78**(3): 1551-60.

Roth, C. M. and A. M. Lenhoff (1995). "Electrostatic and Van-Der-Waals Contributions to Protein Adsorption - Comparison of Theory and Experiment." Langmuir **11**(9): 3500-3509.

Russell, R. B. and G. J. Barton (1992). "Multiple Protein-Sequence Alignment from Tertiary Structure Comparison - Assignment of Global and Residue Confidence Levels." Proteins-Structure Function and Genetics **14**(2): 309-323.

Schellenberger, A. (1998). "Sixty years of thiamin diphosphate biochemistry." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1385**(2): 177-186.

Schrag, J. D., Y. G. Li, S. Wu and M. Cygler (1991). "Ser-His-Glu triad forms the catalytic site of the lipase from Geotrichum candidum." Nature **351**(6329): 761-4.

Seifert, A. and J. Pleiss (2009). "Identification of selectivity-determining residues in cytochrome P450 monooxygenases: A systematic analysis of the substrate recognition site 5." Proteins-Structure Function and Bioinformatics **74**(4): 1028-1035.

Seifert, A., S. Vomund, K. Grohmann, S. Kriening, V. B. Urlacher, et al. (2009). "Rational Design of a Minimal and Highly Enriched CYP102A1 Mutant Library with Improved Regio-, Stereo- and Chemoselectivity." Chembiochem **10**(5): 853-861.

Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden and R. E. Sockett (2005). "Variation in the strength of selected codon usage bias among bacteria." Nucleic Acids Res **33**(4): 1141-53.

Sheehan, D. and R. FitzGerald (1996). "Ion-exchange chromatography." Methods in Molecular Biology **59**: 145-50.

Sheehan, D. and S. O'Sullivan (2001). Ion Exchange Chromatography.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, et al. (2001). "dbSNP: the NCBI database of genetic variation." Nucleic Acids Research **29**(1): 308-311.

Shils, M. E. (2006). Modern Nutrition in Health and Disease (Modern Nutrition in Health & Disease, Lippincott Williams & Wilkins.

Slimko, E. M. and H. A. Lester (2003). "Codon optimization of Caenorhabditis elegans GluCl ion channel genes for mammalian cells dramatically improves expression levels." J Neurosci Methods **124**(1): 75-81.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-8.

Stillger, T., M. Pohl, C. Wandrey and A. Liese (2006). "Reaction engineering of benzaldehyde lyase from Pseudomonas fluorescens catalyzing enantioselective C-C bond formation." Organic Process Research & Development **10**(6): 1172-1177.

Tejo, B. A., A. B. Salleh and J. Pleiss (2004). "Structure and dynamics of Candida rugosa lipase: the role of organic solvent." J Mol Model **10**(5-6): 358-66.

Thanaraj, T. A. and P. Argos (1996). "Protein secondary structural types are differentially coded on messenger RNA." Protein Sci **5**(10): 1973-83.

Thanaraj, T. A. and P. Argos (1996). "Ribosome-mediated translational pause and protein domain organization." Protein Sci **5**(8): 1594-612.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Trodler, P., J. Nieveler, M. Rusnak, R. D. Schmid and J. Pleiss (2008). "Rational design of a new one-step purification strategy for Candida antarctica lipase B by ion-exchange chromatography." Journal of Chromatography A **1179**(2): 161-167.

Varenne, S., J. Buc, R. Lloubes and C. Lazdunski (1984). "Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains." J Mol Biol **180**(3): 549-76.

Vincent, F., S. J. Charnock, K. H. Verschueren, J. P. Turkenburg, D. J. Scott, et al. (2003). "Multifunctional xylooligosaccharide/cephalosporin C deacetylase revealed by the hexameric structure of the Bacillus subtilis enzyme at 1.9A resolution." J Mol Biol **330**(3): 593-606.

Walsh, C. T., S. Garneau-Tsodikova and G. J. Gatto, Jr. (2005). "Protein posttranslational modifications: the chemistry of proteome diversifications." Angew Chem Int Ed Engl **44**(45): 7342-72.

Wang, J. J. L., P. R. Martin and C. K. Singleton (1997). "Aspartate 155 of human transketolase is essential for thiamine diphosphate magnesium binding, and cofactor binding is required for dimer formation." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1341**(2): 165-172.

Wang, J. J. L., P. R. Martin and C. K. Singleton (1997). "A transketolase assembly defect in a Wernicke-Korsakoff syndrome patient." Alcoholism-Clinical and Experimental Research **21**(4): 576-580.

Warshel, A. (1981). "Calculations of enzymatic reactions: calculations of pKa, proton transfer reactions, and general acid catalysis reactions in enzymes." Biochemistry **20**(11): 3167-77.

Warshel, A., G. Naray-Szabo, F. Sussman and J. K. Hwang (1989). "How do serine proteases really work?" Biochemistry **28**(9): 3629-37.

Wei, Y., J. A. Contreras, P. Sheffield, T. Osterlund, U. Derewenda, et al. (1999). "Crystal structure of brefeldin A esterase, a bacterial homolog of the mammalian hormone-sensitive lipase." Nat Struct Biol **6**(4): 340-5.

Widmann, M. (2004). Systematische Analyse und Optimierung expressionsrelevanter Faktoren auf DNA Ebene, Universität Stuttgart.

Widmann, M. and P. Christen (1995). "Differential effects of molecular chaperones on refolding of homologous proteins." FEBS Lett **377**(3): 481-4.

Widmann, M., M. Clairo, J. Dippon and J. Pleiss (2008). "Analysis of the distribution of functionally relevant rare codons." BMC Genomics **9**: 207.

Yadava, A. and C. F. Ockenhouse (2003). "Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems." Infect Immun **71**(9): 4961-9.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics **17**(9): 847-8.

Zeitler, K. (2005). "Extending mechanistic routes in heterazolium catalysis-promising concepts for versatile synthetic methods." Angewandte Chemie-International Edition **44**(46): 7506-7510.

Zhang, S., E. Goldman and G. Zubay (1994). "Clustering of low usage codons and ribosome movement." J Theor Biol **170**(4): 339-54.

Zhao, J. and C. J. Zhong (2009). "A review on research progress of transketolase." Neurosci Bull **25**(2): 94-9.

Zhou, Z., P. Schnake, L. Xiao and A. A. Lal (2004). "Enhanced expression of a recombinant malaria candidate vaccine in Escherichia coli by codon optimization." Protein Expr Purif **34**(1): 87-94.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Verwendung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Stuttgart, April 2010

Michael Widmann