Protein Engineering with Genetic Selection:
Tolerance of Enzyme Activity to Sequence Change

A dissertation submitted to

ETH Zurich

for the degree of
Doctor of Natural Sciences

presented by

Werner Besenmatter

Dipl.-Ing. Technische Universität Wien
born 12.10.1973
citizen of Austria

accepted on the recommendation of
Prof. Dr. Donald Hilvert, examiner
Prof. Dr. Rudolf Glockshuber, co-examiner

Zurich, 2007
“A couple of months in the laboratory can save a couple of hours in the library.”

Frank H. Westheimer, chemistry professor
Parts of this thesis have been published

Besenmatter W., Kast P. and Hilvert D. Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. 
Proteins (2007) 66, 500-506

Besenmatter W., Kast P. and Hilvert D. 
New enzymes from combinatorial library modules. 
# Table of Contents

Acknowledgements ........................................ VIII
Abstract ................................................................ IX
Zusammenfassung .............................................. XIII
Abbreviations .................................................. XVII
  Symbols for Nucleotides ................................ XVIII
  Symbols for Amino Acids ................................ XIX

1. Introduction .................................................. 1
  1.1. Enzyme Engineering ................................. 1
    Enzyme Engineering Methods ...................... 2
    Theoretical Aspects of Enzymes ................. 4
  1.2. Library Design ......................................... 4
    Subsets of Amino Acids ............................. 5
    Quality and Diversity of the Library .......... 7
    Potential Pitfalls ..................................... 8
  1.3. Focus of this Study .................................. 10
    The Enzyme Chorismate Mutase (CM) .......... 10
    Genetic Selection .................................... 12
    Goals of this Study .................................. 14

2. Relative Tolerance of Mesostable and Thermostable Protein Homologs to Extensive Mutation ................................. 15
  2.1. Introduction ......................................... 15
  2.2. Results and Discussion ........................... 16
    Library Design ....................................... 16
    Library Construction and Characterization .... 17
    Sequence Analysis ................................... 20
    Implications for Enzyme Engineering and Design 22

3. Enzyme Engineering by *in silico* Design and *in vivo* Selection with the Minimized Amino Acid Set IKEA .... 25
  3.1. Introduction ......................................... 25
3.2. Results and Discussion

H1 Library Design

In vivo Selection Results

In silico Selection Results

Designer23 Characterization

H2-H3 Libraries

Perspective

4. Plasticity of an Enzyme Active Site

4.1. Introduction

4.2. Results and Discussion

Library Design and Construction

Library Characterization

Analysis of Active Sequences

Characterization of Selected Clones

Conclusions

5. Outlook

6. Materials and Methods

6.1. Reagents and General Procedures

6.2. Vector Construction

6.2.1. Vectors for the Libraries Described in Chapter 2 and 3

6.2.2. Vectors for the Libraries Described in Chapter 3

6.2.3. Vectors for the Libraries Described in Chapter 4

6.3. Library Construction

6.3.1. Synthesis of the FILM-DENK Libraries (Chapter 2)

6.3.2. Synthesis of the IKEA Libraries (Chapter 3)

Synthesis of the IKEA H1 Library (Chapter 3)

Synthesis of the IKEA H2-H3 Libraries (Chapter 3)

6.3.3. One-Pot Synthesis of the Active Site Libraries (Chapter 4)

6.4. Selection for CM Activity

6.5. In vivo CM Activity Assay

6.6. In silico Methods

6.7. Synthesis of Designer23
Acknowledgements

I would like to thank and I am deeply grateful to ...

Prof. Donald Hilvert for becoming my Doktorvater, for the granted freedom, for his constant support, driving enthusiasm, highlighting scientific questions interesting beyond practical applications, the detailed feedback to my texts and his open door.

Prof. Rudolf Glockshuber for readily accepting to be my co-examiner.

Dr. Peter Kast, Dr. Ken Woycechowsky, Dr. Sean Taylor, Dr. Dominik Künzler and Dr. Felix Hausch for many fruitful scientific discussions.

Dr. Alfonso Jaramillo, Prof. Shoshana Wodak and Prof. Martin Karplus for computational prediction of designer23.

my student Ann Babtie for assistance with sequencing, my students Maren Butz, Sabine Seeholzer, Stephanie Schalbetter and Corinne Lochmatter for assistance with retransformations.

Dr. Sean Taylor and Dr. Kai Walter for preliminary experiments with the MjCM library. Dr. Ying Tang for preparation of oligonucleotides with the "split and mix" technique. Dr. Dominik Künzler for chorismate. Severin Sasso for assistance with the CD-spectrometer. Dr. Alexander Kienhöfer and Andreas Kleeb for frequent computer support.

Giulio Casi, Caroline Aldag, Bigna Wörsdörfer, Roger Müller, Dr. Marianne Gamper, Dr. Adrian Hugenmatter, Dr. Florian Seebeck and all other group members already mentioned above, for making my stay so pleasant in the Hilvert-group.

my parents and last but not least Tiina Stämpfli for invaluable support.

Thank you all, it wouldn't have been possible without you!
Abstract

In this thesis the frequency of functional proteins in sequence space is explored. The diversity of possible protein sequences exceeds the number of atoms on our earth by far. Only a miniscule fraction of this theoretically possible diversity is realized in natural proteins — including all enzymes, receptors, channels and pumps — that make life viable. Although these proteins can also be useful outside of their normal biological environment, sequences are not found in nature for every conceivable application of proteins. Therefore, protein engineers are developing a toolbox of methods to create new proteins with tailored properties. One of the most powerful strategies is directed evolution, which entails the selection of desirable DNA sequences from a diversity of many man-made variants — so-called DNA libraries.

Each of the three parts of this thesis investigates the tolerance of enzyme activity to substantial sequence change using a library-based approach. The design of libraries to increase the likelihood of success, specifically to encode the information needed to create an active enzyme, is a central issue. The ability of individual library members to catalyze a chemical reaction provides a functional readout for proper protein folding. The enzyme chorismate mutase served as a model system in our studies, because a potent survival-selection assay for active chorismate mutases is available, enabling the testing of millions of different variants simultaneously. This selection system is based on a chorismate mutase-deficient bacterial strain. Cells of this strain only survive if the culture medium contains the vital amino acids phenylalanine and tyrosine or if a cell can synthesize these nutrients itself due to the provision of a functional chorismate mutase gene. Subsequently, statistical analysis of selected sequences is used to afford insights into the sequence-structure-function relationships of this enzyme.

In the first part of this work (Chapter 2), the influence of the stability of starting structures on the chance of finding active enzymes in randomized libraries was investigated. A thermostable chorismate mutase from the microorganism *Methanococcus jannaschii* was compared to the homologous enzyme from the mesophilic bacterium *Escherichia coli* with respect to their capacity to accommodate extensive mutation. The N-terminal helix comprising about 40% of these proteins was randomized at the genetic level using a binary pattern of hydrophobic and hydrophilic residues based on the
respective wild-type sequences. Catalytically active library members, identified by the genetic selection system, were found approximately 10-times more frequently with the thermostable starting structure compared to its mesostable counterpart. Moreover, detailed sequence analysis revealed that functional \textit{M. jannaschii} enzyme variants contained a smaller number of conserved positions and tolerated greater variability of amino acids at individual sequence positions. These results thus highlight the greater robustness of the thermostable protein with respect to amino acid substitution, while identifying specific sites important for constructing active enzymes. Overall, they support the notion that redesign projects will benefit from using a thermostable starting structure, even at high mutational loads.

In the second part of this work (Chapter 3), the 4-amino-acid-alphabet IKEA was tested experimentally for the assembly of functional enzymes as an alternative to the 20 canonical amino acids, which have redundant side-chain properties. As in Chapter 2, the first helix H1 of \textit{E. coli} chorismate mutase was randomized and catalytically active IKEA variants were identified with the \textit{in vivo} selection system. H1 helical modules constructed from this 4-amino-acid-alphabet afforded functional sequences roughly 10-times more frequently than the 8-amino-acid-alphabet FILM-DENK, which was used in Chapter 2. In parallel to the \textit{in vivo} selection experiments, \textit{in silico} predictions were made using a computational model based on protein folding energies, which confirmed the experimental observations regarding the viability of the IKEA building block set. The highest scoring \textit{in silico} sequence was tested experimentally. The predicted sequence had a high α-helical content and stability similar to the wild-type \textit{E. coli} chorismate mutase. It was also catalytically active \((k_{\text{cat}}/K_m = 1,400 \text{ M}^{-1}\text{s}^{-1})\), even though stability alone was considered in the computation. This success encouraged efforts to replace the other two helices of \textit{E. coli} chorismate mutase with IKEA modules as well, in the hope of generating a fully minimized enzyme. A variety of libraries were constructed and different strategies explored. Unluckily, no active variant — in addition to the active H1 IKEA variants — was found in four different libraries, each containing more than 10 million variants. This surprising result may reflect particularly demanding sequence positions in the third helix. Experiments to test this hypothesis are proposed.

In the third part of this work (Chapter 4), the side-chains that form the active site of the \textit{E. coli} chorismate mutase were randomized. The crystal structure of the enzyme X
shows that these side-chains point from all directions towards the completely buried transition state analog in the active site. Two libraries were constructed. In the first, all 7 polar active site residues were randomized. In the second library, we randomized all 12 active site residues, the 7 polar plus 5 apolar amino acids, whose side-chains contact the bound transition state analog. To minimize experimental effort and resources associated with library assembly, a one-pot synthesis method was developed. From each library, more than 10 million variants were assayed with the \textit{in vivo} selection system. In the large library encompassing the entire active site, one active variant was found, whereas in the other library, which focused on the polar positions, hundreds of active variants were found. Among the 140 sequenced clones, which were very different at the DNA level, 17 unique protein sequences were found with activities of biological significance. Most differences in these sequences are due to two highly variable active site residues at positions 52 and 84. We found all previously published variants, which are theoretically accessible with this library and which have activities above the threshold of the selection system (corresponding to about one thousandth of the wild-type catalytic efficiency or enzyme concentration). Previously unknown and unexpected variants were retransformed to confirm their \textit{in vivo} activities. Five variants were produced and purified to determine their activities \textit{in vitro}. Because only variations of the wild-type active site were found, it can be concluded that alternative active site architectures are either nonexistent or extremely rare — less than one in 10 million in these libraries. Our potent selection-system would have allowed us to capture alternative active site configurations, even if they would have conferred only marginal activities. We conclude that screening or selection experiments with randomized libraries are unlikely to afford enzymes catalyzing new reactions, even under carefully controlled and optimized conditions as applied here. The way forward to create new active sites will presumably require a combination of complementary methods, specifically computational modeling and directed evolution, the two most powerful enzyme engineering approaches used today.

Applying genetic selection to large libraries has allowed the exploration of vast sequence space under defined conditions. Functional sequences, which are extremely rare, could be found with the applied survival-selection system. Nevertheless, exploring sequence space to reveal proteins with completely new activities remains an enormous
challenge, because the sequences isolated and characterized so far present only a minute fraction of all possible sequences and functions. Because most of sequence space lacks function and the islands of function may not be connected directly, it is unlikely that novel activities will be encountered through consecutive rounds of point mutagenesis. To search sequence space more effectively, library design should focus on libraries enriched in folded structures, because structure is a prerequisite to function. Knowledge of three-dimensional structures, basic folding principles, and enzyme mechanism — perhaps complemented with sophisticated computational methods — will help enzyme engineers to design more effective sequence libraries to achieve their goals.
Zusammenfassung


Im ersten Teil dieser Arbeit (Kapitel 2) wurde der Einfluss der Stabilität der Ausgangsstruktur auf die Chancen untersucht, aktive Enzyme in randomisierten Bibliotheken zu finden. Eine thermostabile Chorismatmutase vom Mikroorganismus

Im zweiten Teil dieser Arbeit (Kapitel 3) wurde das 4-Aminosäure-Alphabet IKEA experimentell getestet, für das Zusammensetzen von funktionstüchtigen Enzymen, als eine Alternative zu den 20 kanonischen Aminosäuren, welche redundante Seitenketten Eigenschaften enthalten. Wie im Kapitel 2 wurde die erste Helix H1 von der *E. coli* Chorismatmutase randomisiert und katalytisch aktive IKEA Varianten wurden mit dem *in vivo* Selektions-System identifiziert. H1 Helix-Module, die mit diesem 4-Aminosäure Alphabet geschrieben wurden, gaben ungefähr 10-mal häufiger funktionsfähige Sequenzen als das 8-Aminosäure-Alphabet FILM-DENK, welches in Kapitel 2 benutzt wurde. Parallel zu den *in vivo* Selektions-Experimenten wurden *in silico* Vorhersagen mit einem Computer-Modell, basierend auf Protein-Faltungsenergien, gemacht, welches die experimentellen Beobachtungen der Lebensfähigkeit des IKEA Bausteinsatzes bestätigten. Die höchstbewertete *in silico* Sequenz wurde experimentell getestet. Die vorhergesagte Sequenz hatte einen hohen α-Helix-Gehalt und Stabilität, ähnlicher dem *E. coli* Chorismatmutase Wildtyp. Sie war auch katalytisch aktiv ($k_{\text{cat}}/K_m = 1400 \text{ M}^{-1}\text{s}^{-1}$), obwohl alleine die Stabilität in den Berechnungen berücksichtigt wurde. Dieser Erfolg ermutigte zu Anstrengungen auch die anderen zwei Helices der *E. coli* Chorismatmutase.

alternative Architekturen der aktiven Tasche nicht existieren oder sehr selten sind —
weniger als eine in 10 Millionen in diesen Bibliotheken. Unser potentes Selektions-
System hätte das Aufspüren von alternativen Anordnungen der aktiven Tasche gestattet,
selbst wenn sie nur geringfügige Aktivitäten besäßen. Wir folgern, dass es
unwahrscheinlich ist durch Selektionieren bzw. Durchsuchen von randomisierten
Bibliotheken Enzyme zu finden, die neue Reaktionen katalysieren, selbst unter sorgsam
kontrollierten und optimierten Rahmenbedingungen wie den hier angewandten. Die
Vorwärtsstrategie zur Erschaffung neuer aktiver Taschen wird vermutlich eine
Verbindung von zwei sich ergänzenden Methoden sein, und zwar das Modellieren am
Computer und die gesteuerte Evolution, die zwei leistungsfähigsten Ansätze für Enzym-
Ingenieure, die es heute gibt.

Die Anwendung der genetischen Selektion auf große Bibliotheken erlaubte es, einen
unheimlich weiten Sequenz-Raum zu erkunden. Funktionsfähige Sequenzen, welche
extrem selten vorkommen, sind mit dem verwendeten Überlebens-Selektions-System
auffindbar. Nichtsdestotrotz bleibt die Erforschung des Sequenzraumes, um Proteine mit
komplett neuen Aktivitäten zu entdecken, eine enorme Herausforderung, da jene
Sequenzen, die bereits isoliert und charakterisiert wurden nur einen winzigen Teil aller
möglichen Sequenzen und Funktionen ausmachen. Da der Großteil des Sequenzraumes
ohne Funktionalität ist und Inseln an Funktionalität nicht direkt verbunden sein müssen,
ist es unwahrscheinlich, dass man auf neuartige Aktivitäten durch aufeinander folgende
Zyklen von Punkt-Mutagenese stößt. Um den Sequenzraum effektiver zu durchsuchen,
sollten Bibliotheken entworfen werden, welche angereichert sind an gefalteten
Strukturen, da Struktur eine Voraussetzung für Funktion ist. Kenntnisse von 3D-
Strukturen, grundlegenden Faltungsprinzipien, und Enzymmechanismen — vielleicht
ergänzt durch ausgeklügelte Computermethoden — werden den Enzym-Ingenieuren
helfen effektivere Sequenz-Bibliotheken zu entwerfen um ihre Ziele zu erreichen.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>three-dimensional</td>
</tr>
<tr>
<td>ADO</td>
<td>assembly of designed oligonucleotides</td>
</tr>
<tr>
<td>Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>ampicillin resistance</td>
</tr>
<tr>
<td>BCA</td>
<td>bicinchoninic acid</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>CD</td>
<td>circular dichroism</td>
</tr>
<tr>
<td>CM</td>
<td>chorismate mutase</td>
</tr>
<tr>
<td>Da</td>
<td>Dalton (equivalent to g/mol)</td>
</tr>
<tr>
<td>DHR</td>
<td>degenerate homoduplex recombination</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>EcCM</td>
<td><em>Escherichia coli</em> CM</td>
</tr>
<tr>
<td>FACS</td>
<td>fluorescence activated cell sorting</td>
</tr>
<tr>
<td>H1</td>
<td>the first, N-terminal helix</td>
</tr>
<tr>
<td>H2</td>
<td>the second helix</td>
</tr>
<tr>
<td>H3</td>
<td>the third helix</td>
</tr>
<tr>
<td>IKEA</td>
<td>one letter code of isoleucine, lysine, glutamate and alanine</td>
</tr>
<tr>
<td>IPL</td>
<td>isochorismate pyruvate lyase</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl-β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>ITCHY</td>
<td>iterative truncation for the creation of hybrid enzymes</td>
</tr>
<tr>
<td>LB</td>
<td>Luria Bertani</td>
</tr>
<tr>
<td>min</td>
<td>minute(s)</td>
</tr>
<tr>
<td>MjCM</td>
<td><em>Methanococcus jannaschii</em> CM (not <em>Meloidogyne javanica</em> CM)</td>
</tr>
<tr>
<td>Mw</td>
<td>molecular weight</td>
</tr>
<tr>
<td>MWCO</td>
<td>molecular weight cut off</td>
</tr>
<tr>
<td>NExT</td>
<td>nucleotide exchange and excision technology</td>
</tr>
<tr>
<td>OD&lt;sub&gt;600&lt;/sub&gt;</td>
<td>optical density at 600 nm</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PAGE</td>
<td>polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PaIPL</td>
<td><em>Pseudomonas aeruginosa</em> IPL</td>
</tr>
</tbody>
</table>
PBS  phosphate-buffered saline (10 mM PO_4^{3-}, 160 mM NaCl, pH = 7.5)
PCR  polymerase chain reaction
PMSF  phenylmethanesulfonfyl fluoride
RACHITT  random chimeragenesis on transient templates
RDA  recombination-dependent exponential amplification
SCRATCHY  a DNA recombination method based on ITCHY and DNA shuffling
SDS  sodium dodecyl sulfate
StEP  staggered extension process
TIM  triose phosphate isomerase
TSA  transition state analog
UV  ultraviolet
ε  molar extinction coefficient
Θ  molar ellipticity per residue

Symbols for Nucleotides
A          adenine
C          cytosine
G          guanine
T          thymine
U          uracil
R          A or G         purine
Y          C or T         pyrimidine
S          C or G         aide-mémoire: strong base-pairing
W          A or T         aide-mémoire: weak base-pairing
K          G or T         aide-mémoire: keto
M          A or C         aide-mémoire: amino
B          C or G or T     not A
D          A or G or T     not C
H          A or C or T     not G
V          A or C or G     not T
N          A or C or G or T any nucleotide
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Code</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Ala</td>
<td>alanine</td>
</tr>
<tr>
<td>B</td>
<td>Asx</td>
<td>aspartic acid or asparagine</td>
</tr>
<tr>
<td>C</td>
<td>Cys</td>
<td>cysteine</td>
</tr>
<tr>
<td>D</td>
<td>Asp</td>
<td>aspartic acid</td>
</tr>
<tr>
<td>E</td>
<td>Glu</td>
<td>glutamic acid</td>
</tr>
<tr>
<td>F</td>
<td>Phe</td>
<td>phenylalanine</td>
</tr>
<tr>
<td>G</td>
<td>Gly</td>
<td>glycine</td>
</tr>
<tr>
<td>H</td>
<td>His</td>
<td>histidine</td>
</tr>
<tr>
<td>I</td>
<td>Ile</td>
<td>isoleucine</td>
</tr>
<tr>
<td>J</td>
<td></td>
<td>leucine or isoleucine</td>
</tr>
<tr>
<td>K</td>
<td>Lys</td>
<td>lysine</td>
</tr>
<tr>
<td>L</td>
<td>Leu</td>
<td>leucine</td>
</tr>
<tr>
<td>M</td>
<td>Met</td>
<td>methionine</td>
</tr>
<tr>
<td>N</td>
<td>Asn</td>
<td>asparagine</td>
</tr>
<tr>
<td>P</td>
<td>Pro</td>
<td>proline</td>
</tr>
<tr>
<td>Q</td>
<td>Gln</td>
<td>glutamine</td>
</tr>
<tr>
<td>R</td>
<td>Arg</td>
<td>arginine</td>
</tr>
<tr>
<td>S</td>
<td>Ser</td>
<td>serine</td>
</tr>
<tr>
<td>T</td>
<td>Thr</td>
<td>threonine</td>
</tr>
<tr>
<td>U</td>
<td>Sec</td>
<td>selenocysteine</td>
</tr>
<tr>
<td>V</td>
<td>Val</td>
<td>valine</td>
</tr>
<tr>
<td>W</td>
<td>Trp</td>
<td>tryptophan</td>
</tr>
<tr>
<td>X</td>
<td>Xaa</td>
<td>any amino acid</td>
</tr>
<tr>
<td>Y</td>
<td>Tyr</td>
<td>tyrosine</td>
</tr>
<tr>
<td>Z</td>
<td>Glx</td>
<td>glutamic acid or glutamine</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Enzyme Engineering

Enzymes can accelerate chemical reactions millions to billions of times compared to the uncatalyzed reaction. These catalytically active proteins function in water under mild conditions, such as room temperature and atmospheric pressure, and thus are environmentally benign. An advantage of enzymes — compared to conventional catalysts — is their often high chemo-, regio- and stereospecificity, which leads to less waste and enantiopure products. Natural enzymes catalyze a vast spectrum of reactions, from simple hydrolyses, to C-C bond formations, halogenations,\textsuperscript{1} fixation of nitrogen, and synthesis of large and complex compounds which exert biological activities in minute amounts.

Due to the above properties, enzymes are attractive tools to produce pharmaceuticals, agrochemicals, animal feeds, renewable fuels and chemical intermediates, and also to break down environmental contaminations and stains in textiles. However, natural enzymes — evolved to support the survival and reproduction of their host organisms — have not adapted to applications outside their normal biological context. Due to their natural history, some enzymes have properties undesirable for applications, e.g. insufficient stability or inhibition by small amounts of product or by components of a complex substrate feedstock. Some obstacles can be alleviated by proper implementation, e.g. by improving stability through enzyme immobilization or the use of whole cell catalysts. However, many application challenges are best tackled by changing an enzyme's amino acid sequence, which contains all the information needed for an enzyme's structure and function. This kind of engineering can improve an enzyme's stability or tolerance towards organic solvents, detergents, proteases, oxidation, extremes of pH, pressure and temperature (thermodynamic and kinetic stability, measured as melting temperature and half-life-time, respectively). Furthermore, enzyme engineering can be used to change the pH and temperature optima of an enzyme, to increase catalytic turnover and enantioselectivity, and to broaden the range of accepted substrates. Additionally, protein engineers are attempting to design enzymes that catalyze new reactions, because natural enzyme diversity does not provide for every practical catalysis need. Finally, enzyme engineering can also be used to
investigate academic questions (e.g. to probe an enzyme's reaction mechanism, structure and stability).2,3

Enzyme Engineering Methods

Many different approaches to enzyme engineering have been developed, which build on structural information,4,5 sequence alignments,6,7 computational modeling,8-10 directed evolution11-13 or the immune system.14-16

A promising general approach for the design of new enzymes is built on the concept of transition state stabilization, namely raising antibodies against transition state analogs.14,17 Although catalytic antibodies show high selectivities, their modest catalytic efficiencies have restricted their applications.15

Another hypothesis-driven approach, so-called "rational design" requires high informational input, in the form of a 3D-structure, reaction mechanism or knowledge of active site residues, in addition to intuition, to design new enzymes.4,5,18 Due to a publication bias (successes get published, failures often do not) one might get the impression that most rationally designed mutations are successful predictions and not just serendipity.19,20 However, the main limitation of rational design is that detailed structural information is often not available for the protein of interest.

Compared to the limited number of solved crystal structures, a huge number of sequences is readily available, which can be utilized for enzyme engineering. Sequence alignments of proteins from mesophiles alone can be used to engineer thermostable proteins.6 The strategy involves mutating an amino acid at a given position to the most frequent residue at that position in an alignment of homologous proteins, and then combining individual mutations. The underlying principles of the consensus approach have been rationalized in different ways, either using statistical thermodynamics or based on the hyperthermophilic common ancestor hypothesis.6,7,21 But regardless of the reasons, it is an effective method for optimizing protein production and stability.

A more fundamental approach to protein engineering uses computer models of protein structures. Computational protein design relies on force fields (representing interaction energies), rotamer libraries (representing likely side-chain conformations) and efficient sequence-selection procedures (e.g. dead-end elimination, Monte-Carlo and
genetic algorithms). Current force fields and selection procedures have been sufficiently refined to enable the de novo design of a novel protein fold, although one that lacks a function.\textsuperscript{22} Computational re-design has succeeded in introducing TIM-activity into a catalytically inert structure.\textsuperscript{10} This is a remarkable achievement of computational protein design, even though high-resolution structures of wild-type TIM aided the design,\textsuperscript{23,24} and even though 14 predictions were tested and the best sequence was improved by directed evolution, the resulting protein had a 300-fold lower catalytic efficiency than wild-type TIM.\textsuperscript{10}

In contrast to the engineering approaches described above, directed evolution, and library-based approaches in general, do not require a priori information input.\textsuperscript{11-13} However, if such information is available it can be incorporated in the library design. The concept of directed evolution is based on the Darwinian principles of evolution, which are mutation (diversity), selection (survival of the fittest) and amplification (reproduction). Like natural evolution, multiple traits of a protein can be optimized simultaneously with directed evolution by applying different selection conditions. The selection pressure used for directed evolution can be without precedent in nature, and thus protein properties can be evolved, which are desirable for applications, but detrimental to survival in nature. The numerous mutation methods, which have been developed for directed evolution, include random mutagenesis (e.g. error-prone PCR, mutator strains,\textsuperscript{25} physical and chemical mutagens, like UV-light, nitrous acid and hydroxylamine),\textsuperscript{26} directed mutagenesis (based on synthetic oligonucleotides; Chapter 1.2.) and recombination (either by cleaving and recombining the parent DNA, e.g. DNA shuffling,\textsuperscript{27,28} RACHITT,\textsuperscript{29} NExT,\textsuperscript{30} ITCHY\textsuperscript{31} and SCRATCHY;\textsuperscript{32} or oligonucleotide based, e.g. synthetic shuffling,\textsuperscript{33} ADO\textsuperscript{34} and DHR,\textsuperscript{35} or PCR-based, e.g. StEP\textsuperscript{36} and RDA-PCR,\textsuperscript{37} or in vivo recombination).\textsuperscript{38-41} The invented acronyms for recombination methods (e.g. ITCHY and SCRATCHY) are almost as vivid as the names of genes from fruit flies (e.g. ken and barbie).\textsuperscript{31,32,42} The identification of improved variants from the library of mutants can be attained by screening (from manual testing to high-throughput screening), sorting (e.g. FACS), panning (for binders), and selection (of proteins that confer antibiotic-resistance or enable the synthesis of an essential nutrient).\textsuperscript{43,44} Furthermore, for directed evolution experiments the linkage between genotype and phenotype is crucial and can be achieved with various methods (in vivo,
in vitro compartmentalization by water-in-oil emulsions, phage display, mRNA display, ribosome display, cell-surface display, yeast n-hybrid systems, and spatially in microtiter plates or on protein chips)\textsuperscript{43,44}

Theoretical Aspects of Enzymes

Beside the practical aspects of enzyme engineering described above, theoretical questions concerning enzymes are also fascinating. For example, how natural enzymes achieve immense rate accelerations compared to the corresponding uncatalyzed reactions is still debated.\textsuperscript{45} These rate accelerations are generally rationalized as stabilization of the transition state relative to the enzyme-bound ground state.\textsuperscript{45-48} However, the restriction of the substrate to a specific conformation (a strained conformation or a "near attack conformation") may also play a role.\textsuperscript{49,50} A better understanding of these issues may help to design new enzymes with catalytic constants comparable to those of natural enzymes.

1.2. Library Design

In combinatorial approaches to protein engineering, large libraries are searched for variants with desired properties. However, an exhaustive search of all variants of a small protein, 100 amino acids long, would necessitate the generation of $2^{100} \approx 10^{30}$ variants. Even if only one molecule of each variant would be produced, such a library would be massively heavier ($2^{100} \times 110 \text{ g/mol} \times 100 / (6 \times 10^{23} \text{ mol}^{-1}) \approx 2 \times 10^{110} \text{ g}$) than our earth ($6 \times 10^{27} \text{ g}$) and is thus experimentally inaccessible. Even a partial search would be fruitless, unless proteins with desirable properties occur at very high frequencies. To increase the likelihood of finding functional proteins, libraries can be biased towards specific folds, because proper folding is a necessary prerequisite for protein function. For example, one can start from a sequence, which is known to fold into a well-defined 3D-structure, and randomize only the active site residues (Chapter 4); or one can utilize basic structural information, namely a binary pattern of hydrophilic and hydrophobic residues, for library design (Chapters 2 and 3).\textsuperscript{51,52} Protein engineering by binary patterning is based on the principle that hydrophilic residues prefer aqueous surroundings, whereas
A suitable sequence of hydrophilic and hydrophobic residues can direct a polypeptide chain to fold into amphiphilic secondary structures that anneal together to form a defined tertiary structure. The sequence of polar and apolar residues is the major determinant of secondary structure, which can outweigh the propensities inherent in individual amino acids to adopt specific secondary structures. Binary patterning utilizes the intrinsic periodicities of secondary structures (Figure 1), while allowing the identity of the residue at each randomized position to be combinatorially changed.

Figure 1. Simple binary patterns for β-strands (A and C) and α-helices (B and D, respectively). With the patterns shown in (A) and (B), hydrophilic (blue) and hydrophobic (red) residues are segregated to opposite sides of the secondary structures, represented as an arrow for a β-strand (C) and as an α-helical wheel (D). The hydrophilic side is designed to face the aqueous exterior, while the hydrophobic side segregates the apolar interior of the protein.

Subsets of Amino Acids

Another approach to reduce theoretical library sizes is to allow only a subset of the 20 canonical amino acids in the design. The natural set of proteinogenic amino acids contains redundant side-chain properties (2 × -COOH, 2 × -CONH₂, 2 × -OH, 2 × -CH(CH₃)₂). That this redundant amino acid set continues to exist can be considered...
a "frozen accident". In fact, just 9 amino acids (FILMDENKR) are sufficient to construct a protein with chorismate mutase-activity of biological significance. This set of 9 amino acids was chosen because it can be easily encoded at the DNA level. Conceivably, even smaller sets or other combinations of amino acids may be better suited for enzyme engineering. For example, the 4 amino acid set IKEA is one attractive possibility (see Chapter 3), not because of the Swedish company selling minimalistic components that can be assembled into functional furniture, but because this set covers most basic features of amino acids in proteins, including a large and a small hydrophobic residue (Ile and Ala) as well as an acid and a base (Glu and Lys).

Applying binary patterning, the polar amino acids can be easily encoded with degenerate oligonucleotides synthesized from a mixture of A plus G and pure A. The degenerate codon (A+G)AA encodes only Lys (AAA) and Glu (GAA). However, the degenerate codon (A+G)(T+C)T for the apolar amino acids Ile (ATT) and Ala (GCT) additionally encodes the unwanted amino acids Val (GTT) and Thr (ACT). To circumvent the latter problem, many different strategies are possible. The most straightforward approach is to synthesis different oligonucleotides and to mix them before library assembly. However, this strategy is only feasible when relatively few positions are targeted for randomization, because the number of necessary oligonucleotides increases exponentially with the number of targeted positions.

Another solution is to use mixtures of trinucleotides corresponding to the desired codons during DNA synthesis. The necessary trinucleotide phosphoramidites are difficult and expensive to prepare, but have recently become commercially available (from Glen Research, Sterling, USA). More readily prepared trinucleotide phosphoramidites with orthogonal protecting groups have been used in a strategy of parallel codon synthesis. A strategy that does not require special reagents is the "split and mix" technique. In this approach, the resin is split at a targeted position, and then the desired codons are synthesized separately. The resin is then remixed and the synthesis continued. The number of codons synthesized in parallel can be varied at will, in contrast to the binomial strategy with orthogonally protected trinucleotides.

Strategies that do not intervene in DNA synthesis have also been developed. Only conventional primers are required in a strategy called MAX. It is based on the annealing of specific primers on a template that contains the completely randomized
codon NNN at the targeted positions. The specific primers select a specific subset of desired codons. After annealing to the template, specific primers covering all parts of the template are ligated together. The ligated strand can be selectively PCR-amplified with two outside primers that anneal on the overhanging parts of the ligated strand, and thus cannot amplify the NNN template. A sizeable number of primers are needed for this strategy, but the number of oligonucleotides increases only linearly — and not exponentially — with the number of targeted positions. One limitation of the MAX strategy is that it can only be applied if the targeted positions are not too close to each other on the template. A completely different strategy, called SlonoMax (commercialized by Sloning BioTechnology GmbH, Puchheim, Germany), is based on the stepwise ligation of double-stranded triplet building blocks. Mixtures of building blocks allow the synthesis of randomized DNA double-strands.

These strategies release us from the constraints imposed by the genetic code (Appendix 7.2.). They enable the construction of DNA libraries that encode any subset of amino acids. Even if all 20 amino acids should be encoded at one position, the above described strategies have the advantage of removing codon bias associated with the up to six-fold difference in the number of codons for individual amino acids.

**Quality and Diversity of the Library**

To check for bias at the randomized positions and to confirm that the synthesized library conforms to the intended design, a representative number of library members should be sequenced. It is not uncommon to find that only a minority of the library members (often only 30% to 40%) has the correct length. In fact, all types of point mutations — insertions, deletions and substitutions — can be observed in library members that are not subject to selection pressure. These may arise during chemical synthesis of the oligonucleotides, *in vitro* synthesis by a DNA polymerase, or *in vivo* DNA synthesis in *E. coli*. From the observation that such unplanned mutations are generally much more frequent in the stretches derived from synthetic oligonucleotides, we can single out chemical synthesis as the main source of such errors.

Indeed, unplanned mutations are commonly observed in any approach that relies on chemically synthesized oligonucleotides. Single base deletions can originate
during chemical synthesis from incomplete coupling and capping, as well as from inefficient removal of the 5’ protecting group prior to coupling. Insertions can result from cleavage of the 5’ protecting group on the phosphoramidite building blocks before and during coupling. Contamination of the phosphoramidites, potentially due to the use of common supply lines in the DNA synthesizer, can lead to base substitutions. Longer oligonucleotides are more extensively damaged than shorter ones, presumably because of longer exposure to the reagents during synthesis (e.g. non-specific acylations, etc.) and the lower efficiency of purification of correct length oligonucleotides from oligonucleotides just one or a few bases shorter.

Nevertheless, mutations introduced into DNA libraries through oligonucleotide synthesis generally do not represent a serious problem if library members are subsequently subjected to a selection procedure. For instance, the requirement for function automatically eliminates frameshifts leading to dysfunctional gene products.

Potential Pitfalls

Beside the potential problems associated with oligonucleotide synthesis (described in the previous section), it is also useful to reflect about other potential pitfalls and how to circumvent them, before engaging in the construction of large DNA libraries, which generally represents a considerable investment of time and resources.

Small Library Sizes: Many factors could be responsible for a low yield of library clones. Some of these include the use of inappropriate or inefficient restriction endonucleases or competent cells. Contamination by nucleases or by inhibitors of the enzymes employed constitutes another problem, which may require more thorough DNA purification. On the other hand, significant loss of DNA may occur during the numerous purification steps and it is therefore important to monitor the progress of library construction by analytical agarose gels. In particular, smaller fragments are more difficult to handle than longer DNAs. A reduced library size may also result from DNA damage. At any stage, damage can occur, but particularly critical is exposure to UV light after gel electrophoresis. UV-radiation-induced lesions in synthetic oligonucleotides or double-stranded DNA may block subsequent replication in the commonly used RecA-deficient E. coli strains.
**False Positives:** One possible reason for false positives is contamination with wild-type sequence, either from an external sources or from a precursor to the random library. To minimize the risk of external contamination, only new plastic-ware should come in contact with DNA solutions, and filter-tips should be used to pipette DNA solutions. To reduce the risk of contamination from a precursor of the library acceptor, an acceptor plasmid with a stuffer fragment can be used. The stuffer fragment replaces the part of the wild-type sequence that is ultimately replaced with the library module. The length of the stuffer DNA (ca. 1 kb fragments from phage \(\lambda\) or pKSS\(^{66}\) are convenient) should be chosen such that the desired doubly cut acceptor fragment can be easily purified from the singly cut plasmid and that the correct recombinants can later be distinguished from the acceptor vector by a simple analytical restriction digest. A strategy to avoid contamination with wild-type PCR-templates is to use templates that have some of their thymines replaced by uracil. The incorporation of uracil, as well as the subsequent cleavage of uracil-containing DNA can be easily done with different methods *in vitro* and *in vivo*.\(^{30,67}\)

Another possible source of false positives is the co-expression of an active and an inactive library gene in one cell. Addition of salt, which stabilizes double-stranded DNA,\(^{68}\) before incubation at increased temperatures can alleviate this problem. In the absence of salt, short double-stranded library DNA might easily denature upon heating. After cooling, individual DNA strands that are not perfectly complementary could reanneal and form heteroduplex DNA, which might be cloned into the acceptor vector. This DNA would be transformed as one plasmid, but would subsequently be propagated as two plasmid variants in one cell. Optimal protocols therefore also limit exposure of the library modules to chaotropic salts, which are generally employed in gel extraction and DNA purification kits but denature short double-stranded DNA, until the library module is safely ligated to the acceptor plasmid.

Finally, to avoid wasting time on false positives, the match between phenotype and genotype needs to be confirmed. Plasmids isolated from selected library clones are retransformed into the selection strain to verify *in vivo* activity. However, only small DNA amounts should be retransformed, because at high DNA concentrations it could be possible that two different library plasmids will be electroporated into a single cell.
1.3. **Focus of this Study**

This thesis investigates the relation between protein sequence and enzyme activity with library-based approaches. Library design, or how to encode the information needed for catalysis into protein sequences, constituted the starting point for each project. A powerful survival-selection assay for active enzymes allowed probing millions of different variants simultaneously. Statistical analysis of selected sequences gave insight into the sequence-structure-function relationship. In complementary strategies, either structurally important residues were changed and active site residues were kept constant, or the active site residues were modified while keeping the other residues constant.

**The Enzyme Chorismate Mutase (CM)**

In all projects of this thesis, CM was used as a straightforward model for enzyme engineering, because it catalyzes a simple one substrate to one product reaction without the need for cofactors or enzyme bound intermediates (Figure 2).

![Figure 2. Chorismate mutase catalyses the Claisen-rearrangement of chorismate (1) to prephenate (3). In solution, chorismate is in equilibrium between the pseudo-equatorial (1 eq.) and the pseudo-axial conformer (1 ax.). This reaction proceeds via a transition state (2).](image)
Figure 3. Chorismate is central in biosynthetic pathways to aromatic compounds. CM catalyzes the first committed step in the pathway to the aromatic amino acids phenylalanine and tyrosine. Reactions with more than one intermediate are represented by double arrows.

CM (EC 5.4.99.5) catalyzes the last step in the shikimate biosynthetic pathway (Figure 3), before branching to the vital amino acids phenylalanine and tyrosine in bacteria, fungi, plants and nematodes. Because humans and other mammals lack a CM (they need to take up phenylalanine and tyrosine through their diet), this enzyme is a potential target for antibiotics, fungicides and herbicides.

Many CM sequences from different species have been identified, either by direct isolation, or by sequencing of genomes and comparison with known CM genes. CMs have been found to be monofunctional (e.g. from Bacillus subtilis, Thermus thermophilus, Methanococcus jannaschii, Saccharomyces cerevisiae and Aspergillus nidulans) or bifunctional (e.g. from Escherichia coli, Erwinia herbicola, Buchnera aphidicola and Xanthomonas campestris). The bifunctional CMs from E. coli are fused to prephenate dehydratase or prephenate dehydrogenase and thus catalyze consecutive step in the pathway to phenylalanine and tyrosine. Most CMs are located in the cytoplasm, but CMs in the periplasm (from E. herbicola, Salmonella typhimurium and Pseudomonas aeruginosa) and CMs secreted into the extracellular medium (from
Mycobacterium tuberculosis) were found. CMs are also structurally diverse and can be classified as AroQ and AroH. The AroH class CMs (e.g. from *B. subtilis* and *T. thermophilus*) are homotrimeric proteins that form pseudo-α,β-barrels with three active sites. The prototype of the AroQ class CMs from *E. coli* (EcCM) is an all α-helical homodimer with two active sites, which are completely covered by the protein (Figure 4). Most CMs discovered so far belong to the AroQ class, including the CMs from *M. jannaschii*, *M. tuberculosis* and *S. cerevisiae*. A different structural class is the man-made catalytic antibody with CM-activity. It has a standard, all β immunoglobulin structure.

For this study EcCM was used as the principle model protein, because it is the prototype of the most abundant class of CM. Furthermore EcCM's small gene size, encoding 97 amino acids, and its solved crystal structure, make it an excellent target for library design and subsequent genetic selection.

**Figure 4. Structure of EcCM.** One of the two symmetrical subunits is drawn in purple. The long N-terminal helices (H1 and H1') span the homodimer. The transition state analog (green) is completely buried. The figure was prepared with the program MOLMOL.

**Genetic Selection**

The availability of a powerful genetic selection system for CM activity (Figure 5) facilitated the identification of active enzymes in large libraries.
Figure 5. The survival-selection assay used in this study: From library modules to functional enzymes. First, a library module is inserted into the acceptor plasmid by replacing a stuffer fragment. The resulting plasmid library is transformed into the CM-deficient selection strain KA12/pKIMP-UAUC. After transformation, the pool of cells is divided. The lion's share is plated onto M9c minimal medium plates to select for active CMs from the library. CM, which catalyzes the reaction of chorismate to prephenate in the biochemical pathway to phenylalanine (F) and tyrosine (Y), is essential for survival of the strain on minimal medium lacking F and Y. A small aliquot of the cells is plated onto medium that contains F and Y to determine the number of transformants.
CM activity is vital for cells on minimal medium lacking phenylalanine and tyrosine. The CM-deficient *E. coli* cells used for selection can be rescued either by supplying phenylalanine and tyrosine through the culture medium, or by providing the cells with a plasmid that contains a functional CM gene. The selection strain lacks both chromosomal bifunctional CM genes, *pheA* and *tyrA*, encoding the fusion proteins CM•prephenate dehydratase and CM•prephenate dehydrogenase, respectively. To restore the essential activities of the fusion partner, the monofunctional prephenate dehydratase gene *pheC* from *Pseudomonas aeruginosa* and the monofunctional prephenate dehydrogenase gene *tyrA* from *Erwinia herbicola* is supplied with the auxiliary plasmid pKIMP-UAUC. The auxotrophic strain can be rescued if it is transformed with a functional CM gene. To enable the detection of low activity, library members are expressed at a high level (on a high-copy-number vector and transcribed from the strong, constitutive trc promoter). The identification of functional clones among millions of library members is straightforward, because dysfunctional sequences do not appear. With a survival-selection assay larger libraries are experimentally accessible than with high throughput screening. The limiting step is not anymore the assessment of individual library members. Instead, transformation efficiency becomes the bottle neck.

**Goals of this Study**

The general goal of this thesis was to investigate the frequency of functional proteins in sequence space. First, the influence of protein stability on the outcome of directed evolution was examined (Chapter 2). Second, a quest for the minimal set of amino acids sufficient to construct functional proteins was undertaken to reduce the redundancy in the canonical set of 20 amino acids (Chapter 3). Finally, the active site of EcCM was randomized to investigate the chances of finding alternative active sites (Chapter 4). These experiments provide valuable insight into the sequence-structure-function-activity relationship and into the topology of sequence space.
2. Relative Tolerance of Mesostable and Thermostable Protein Homologs to Extensive Mutation

2.1. Introduction

Proteins from natural sources, optimized through evolution and comprising a wealth of distinctive architectures, represent potentially attractive starting materials for the creation of novel receptors, catalysts and advanced materials for diverse applications. These biological macromolecules can be systematically improved or altered through site-directed mutagenesis, guided either by structure,5,6 sequence alignments,6,7 or computational modeling.8-10 Increasingly, efforts to tailor protein properties are also profiting from the tools of directed evolution, which couples random mutagenesis with powerful screening or selection strategies.11-13

Since homologous proteins often differ greatly in their physical properties, the choice of the starting template can be crucial for the success of a protein engineering project. The large difference in stability of orthologs isolated from mesophiles and thermophiles is a case in point. Because of their potential to accommodate a greater number of destabilizing mutations, thermostable proteins have often been preferred as starting points for new designs.95-97 However, such proteins can be more rigid than their mesostable counterparts, and they are sometimes substantially less active at room temperature.98-100 Consequently, one can ask whether thermostable proteins really are better design templates. And if they are, how much better? To answer such questions objectively, systematic comparisons of related proteins of differing stability with respect to their robustness to amino acid substitution are needed. Experimental evidence in support of the notion that enhanced stability confers mutational robustness and hence evolvability has recently been obtained,101,102 but the mutational loads were relatively modest.

In the current study, we have examined the ability of homologous mesostable and thermostable AroQ chorismate mutases to tolerate extensive mutation. Large segments of the genes encoding these enzymes were randomized and subjected to in vivo selection to identify catalytically active variants. Analysis of the respective complementation frequencies and the specific amino acid preferences at individual positions provides direct insight into the relative fitness of the parent proteins.
2.2. Results and Discussion

Library Design

To investigate the relationship between protein stability and tolerance to amino acid substitution, we chose the thermostable chorismate mutase (CM) from *Methanococcus jannaschii* (MjCM)\(^75\) and the homologous CM domain from the bifunctional *E. coli* chorismate mutase-prephenate dehydratase (EcCM)\(^85,103\). These enzymes are about 100 amino acids long and share 31% sequence identity. They both adopt a homodimeric α-helical fold, and exhibit comparable catalytic properties, thermodynamic activation parameters, and inhibition profiles.\(^75\) However, they differ greatly in stability. The enzyme from the thermophile has a 25°C higher \(T_m\) value, and chemical denaturation experiments suggest that it is approximately 5 kcal/mol more stable than EcCM.\(^75\)

The long N-terminal helix, which spans the homodimer and constitutes about 40% of the entire protein, was replaced with modules of randomized sequence in both MjCM and EcCM. Aside from five residues that were held constant either because they are part of the active site or were needed for construction purposes, each position in the helix was mutagenized according to the polarity of the corresponding amino acid in the parent CM (Figure 6).

<table>
<thead>
<tr>
<th>wild-type EcCM</th>
<th>MTSENPLLALREKISALDEKLLALLAERRELAVEVGKAKLLS...</th>
</tr>
</thead>
<tbody>
<tr>
<td>EcCM library</td>
<td>M●●●●●●●●●R●●●●●●●KL●●●●●●●R●●●●●●●●●K●●●...</td>
</tr>
<tr>
<td>wild-type MjCM</td>
<td>MIEKLAERIRKKIDEIDNKILKIAERNSLAKDVAEIKNQL...</td>
</tr>
<tr>
<td>MjCM library</td>
<td>M●●●●●●●●●R●●●●●●●KL●●●●●●●R●●●●●●●●●K●●●...</td>
</tr>
</tbody>
</table>

**Figure 6. Library design.** Binary patterned helical modules were constructed to replace the N-terminal helix of *E. coli* chorismate mutases (EcCM) and its thermostable homolog from *M. jannaschii* (MjCM) based on the distribution of hydrophilic and hydrophobic residues in the corresponding wild-type sequences. Blue and red circles denote mixtures of the four apolar amino acids (Leu, Ile, Met, Phe) and the four polar amino acids (Lys, Glu, Asp, Asn), respectively. Catalytic residues and residues held constant for library construction are specified. Conventional EcCM sequence numbering was used throughout.
Hydrophobic residues were replaced by mixtures of the four apolar amino acids leucine (Leu), isoleucine (Ile), methionine (Met) and phenylalanine (Phe), whereas hydrophilic residues were replaced by mixtures of the four polar amino acids lysine (Lys), glutamate (Glu), aspartate (Asp) and asparagine (Asn). This set of eight building blocks is readily encoded at the DNA level and has been used successfully in previous experiments to replace all the helices in MjCM with binary-patterned modules. The ratio of hydrophobic to hydrophilic residues targeted for mutagenesis is approximately 1:1 for both templates.

**Library Construction and Characterization**

The MjCM and EcCM gene libraries were each constructed from two long oligonucleotides and electroporated into a CM-deficient *E. coli* strain to give $3 \times 10^7$ and $5 \times 10^6$ transformants, respectively (Table I).

<table>
<thead>
<tr>
<th>Library</th>
<th>Library size</th>
<th>Fraction of correct genes</th>
<th>Total number of complementing colonies</th>
<th>Frequency of complementing colonies</th>
</tr>
</thead>
<tbody>
<tr>
<td>EcCM</td>
<td>$5 \times 10^6$</td>
<td>$\approx 37%$</td>
<td>36</td>
<td>approx. 1 in 50,000</td>
</tr>
<tr>
<td>MjCM</td>
<td>$3 \times 10^7$</td>
<td>$\approx 30%$</td>
<td>1980</td>
<td>approx. 1 in 4,500</td>
</tr>
</tbody>
</table>

*a* Number of independent transformants present in the library as determined from two different dilutions of cell suspensions plated under non-selective conditions; the estimated standard error of the mean is 10%.

*b* Derived from 27 (EcCM library) and 68 (MjCM library) sequenced clones of randomly picked library members.

Prior to selection, between 30 and 40% of the library members had the correct length and contained no frameshifts or promoter mutations (Table I), consistent with
results obtained with other libraries constructed from long oligonucleotides. A $\chi^2$ (P < 0.01, df = 5) goodness of fit test of genuine library sequences found no significant deviation from design (Table II), thus confirming that the libraries were unbiased.

### Table II. Codon usage at all randomized positions prior to selection.

<table>
<thead>
<tr>
<th>Codons</th>
<th>EcCM library $^a$</th>
<th>MjCM library $^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>expected</td>
<td>found</td>
</tr>
<tr>
<td><strong>HTS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATG = Met</td>
<td>30 (17%)</td>
<td>34 (19%)</td>
</tr>
<tr>
<td>ATC = Ile</td>
<td>30 (17%)</td>
<td>35 (19%)</td>
</tr>
<tr>
<td>CTG = Leu</td>
<td>30 (17%)</td>
<td>32 (18%)</td>
</tr>
<tr>
<td>CTC = Leu</td>
<td>30 (17%)</td>
<td>28 (16%)</td>
</tr>
<tr>
<td>TTG = Leu</td>
<td>30 (17%)</td>
<td>28 (16%)</td>
</tr>
<tr>
<td>TTC = Phe</td>
<td>30 (17%)</td>
<td>23 (13%)</td>
</tr>
<tr>
<td><strong>RAV</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAG = Lys</td>
<td>30 (17%)</td>
<td>25 (14%)</td>
</tr>
<tr>
<td>AAA = Lys</td>
<td>30 (17%)</td>
<td>30 (17%)</td>
</tr>
<tr>
<td>AAC = Asn</td>
<td>30 (17%)</td>
<td>36 (20%)</td>
</tr>
<tr>
<td>GAG = Glu</td>
<td>30 (17%)</td>
<td>25 (14%)</td>
</tr>
<tr>
<td>GAA = Glu</td>
<td>30 (17%)</td>
<td>29 (16%)</td>
</tr>
<tr>
<td>GAC = Asp</td>
<td>30 (17%)</td>
<td>35 (19%)</td>
</tr>
</tbody>
</table>

$a$ n = 10 EcCM library members analyzed  

$b$ n = 19 MjCM library members analyzed  

Active catalysts were subsequently identified by their ability to complement the CM-deficient host strain under auxotrophic growth conditions. The MjCM and the EcCM libraries yielded 1980 and 36 complementing colonies, respectively, corresponding to a complementation frequency of approximately 0.022% and 0.002% of correct library members (Table I). Thus, functional library members were approximately 10 times more abundant when the thermostable protein was employed.
Figure 7. Sequence analysis of the randomized positions in functional enzymes selected from the EcCM (A) and MjCM (B) libraries. The tables give the numbers of amino acids observed at each position, the bar diagrams above show the calculated percentages of each amino acid at each randomized position. The designed amino acid distribution before selection is shown on the right. Positions with significant deviation...
from design, according to a $\chi^2 (P < 0.01, \text{df} = 3)$ goodness of fit test, are marked by an asterisk. The diagram below the table shows for each amino acid the normalized deviation from design, which was calculated from $(O - E)/N$, where $O$ is the observed number of amino acids, $E$ the expected number of amino acids prior to selection based on the library design, and $N$ the total number of analyzed sequences. $N$ equals 26 and 20 for the active chorismate mutase sequences analyzed from the EcCM and the MjCM libraries, respectively.

**Sequence Analysis**

*Figure 7* summarizes the amino acid distributions found at the randomized positions of 20 active MjCM and 26 active EcCM variants. The results confirm that the N-terminal helix of both proteins is highly tolerant to mutation. Significant deviations from a random distribution (marked by an asterisk) are only observed at roughly one third of the randomized positions. Notably, the MjCM template has fewer such sites (10) than EcCM (13), and the constrained positions in the former are largely a subset of the latter. Detailed analysis of the observed sequence preferences further reveals that hydrophobic residues are more likely to be constrained (40 - 50% of the randomized apolar sites) than hydrophilic residues (17% of the randomized polar sites), although several of the latter are among the sites showing the strongest selection preferences overall. Furthermore, at a given constrained position, the same amino acid tends to be selected in both libraries, but the preference for the selected amino acid is typically less stringent in MjCM than in EcCM, and the range of acceptable amino acids is usually greater. The examples summarized below illustrate these general trends.

Given the importance of hydrophobic residues in packing, it is not surprising that a disproportionate number of the targeted apolar sites show statistically meaningful deviations from a random distribution. The long N-terminal helix of AroQ enzymes forms an antiparallel coiled-coil with its counterpart from the second polypeptide of the homodimer; it also engages in extensive interactions with the other helices in the protein that were held constant by design. As shown in *Figure 7*, six apolar sites in the MjCM library were subject to significant selection pressure (positions 14, 25, 32, 35, 36 and 38), and nine in the EcCM library (the MjCM set, plus residues 17, 22 and 33). They
include residues involved in coiled-coil packing (14, 17, 35, 38) as well as interactions between the N-terminal helix and the rest of the protein (22, 25, 32, 33, 36). Residues 14 and 35 also provide a portion of the hydrophobic surface of the substrate binding pocket, which may have further limited their tolerance to substitution.

At none of the selected apolar sites is an absolute preference for a single amino acid evident, although the aliphatic building blocks leucine and isoleucine are generally favored and the more flexible methionine and the bulky aromatic phenylalanine tend to be statistically underrepresented. The trends observed at positions 14 and 32 are typical. At both sites, 88% of the functional EcCM variants contain isoleucine, whereas the preference for this residue is much less pronounced in MjCM (only 40 - 60% of the active clones). Occasionally, the influence of the original template protein is strongly manifest, leading to divergent choices for the two libraries. For example, at position 25 the respective wild-type residue is strongly favored — leucine in EcCM (88%) and isoleucine in MjCM (75%), presumably because it packs better against the portion of the template protein that was held constant. In a few instances, the wild-type usage is clearly overridden. For instance, a relatively bulky phenylalanine, which is otherwise underrepresented, is selected at position 36 in 35 - 40% of the active MjCM and EcCM clones, even though residue 36 is typically a small amino acid in AroQ proteins (alanine in MjCM and glycine in EcCM). The ability of the protein scaffold to adjust to this large perturbation in volume points to its inherent plasticity.

The relative tolerance of hydrophobic residues to substitution is in accord with their general solvent exposure on the surface of the protein. The exceptions to this general rule are found at positions 15, 18, 29, and 42 in the EcCM library and at positions 15, 18, 30, and 40 in the MjCM library. Although the preferences observed at positions 30, 40 and 42 are modest and not readily rationalized, the strong selection at the other positions can be explained in structural terms. For example at positions 15 and 18, a significant preference for Asp was found in both libraries at both positions, but the requirement for Asp was more stringent in the EcCM than in the MjCM library. In the EcCM and the MjCM library, 81% and 75%, respectively, of functional variants contained Asp at position 15. Noteworthy, the same trend was observed again at position 18. All twenty-six (100%) analyzed active EcCM library members contained Asp at position 18. In contrast, just 85% of the twenty analyzed MjCM variants had Asp at position 18, while 15% contained the isosteric Asn. The aspartates at positions 15 and 18 are second shell
residues that can form salt bridges with the essential active site residues Arg51’ and Arg28’, and thereby help preorganize the binding pocket for catalysis.\textsuperscript{60,85}

The lysine selected at position 29 in all active EcCM variants replaces an arginine in the wild-type protein, which interacts electrostatically with two negatively charged residues, Asp83 of the same polypeptide and Asp69’ from the other subunit. Selection of a positively charged residue at this position allows reconstitution of wild-type-like interactions, which presumably contribute to the stability of the fold. An analogous interaction is not found (or apparently required) in MjCM library members, perhaps because this template has a histidine at position 83 rather than an aspartate.\textsuperscript{107}

**Implications for Enzyme Engineering and Design**

Because the information encoded at the level of primary sequence is highly redundant,\textsuperscript{51,108} a substantial fraction of the amino acids in any given protein can be replaced without disturbing overall structure or biological function.\textsuperscript{51,109-113} This resilience to mutation provides protein designers with the means to tailor existing scaffolds to their needs, as new functions can be accessed without necessarily sacrificing structural integrity. Highly stable proteins from thermophilic organisms, which can be expected to accommodate multiple mutations, represent intuitively attractive starting points for such endeavors. In fact, the choice of a thermostable variant has led to successful designs,\textsuperscript{95-97} although the precise benefit over using an analogous mesostable protein is generally unknown.

The potential advantage of stable templates for protein engineering has been recently addressed experimentally and theoretically.\textsuperscript{101,102,114} Gene-wide mutagenesis studies at low mutation rates have shown that the fraction of functional proteins generally decreases exponentially with the number of random substitutions,\textsuperscript{101,111-113,115} although the decline depends to some extent on the starting fold.\textsuperscript{101,111} This behavior has been successfully reproduced by a lattice protein model, which predicts that the sensitivity to mutation depends both on the structure and stability of the starting protein.\textsuperscript{101} In fact, the specific prediction that stable proteins will afford a higher fraction of folded mutants was found to agree quantitatively with the experimental effects of a few random substitutions on several different starting proteins.\textsuperscript{101,102} As the number of mutations increases,
however, the influence of the original protein should diminish greatly and, ultimately, disappear.

We have examined the benefits of template stability using a fundamentally different approach. Rather than investigating protein tolerance to a small number of mutations, we replaced an entire helical segment in the AroQ scaffold with a module of random sequence. Although the mutational load in our system is thus extremely high, the probability of isolating functional proteins is maintained in an experimentally accessible range by exploiting basic structural information, namely the binary pattern of polar and apolar residues in the randomized module, for library design. Indeed, we find that a substantial fraction of our libraries is functional in the in vivo selection assay — ranging between 1 in ca. 5,000 to 1 in ca. 50,000 sequences, despite the fact that 40% of the entire protein was mutated and only a restricted set of four polar and four apolar building blocks was allowed at each randomized position. Functional variants in both libraries contain between 20 and 29 simultaneous amino acid substitutions, corresponding to approximately one quarter of all residues in each protein. The ability to move so far from the starting sequence without loss of biological function underscores the degeneracy of the message specifying proper folding, dimer assembly, and catalysis in this system.

Contrary to the expectation that the influence of starting protein stability should wane at high mutational loads, a decided advantage is observed for the thermostable MjCM in our experiments relative to the less stable EcCM. This advantage is manifest in (1) a 10-fold higher complementation frequency, (2) fewer “hot spots,” or positions showing significant preference for a particular amino acid, and (3) considerably higher sequence variability at these sites compared to the analogous hot spots in EcCM. Interestingly, and relevant to design and engineering considerations, the constrained sites, which largely overlap for the two proteins, pinpoint critical features in the scaffold that would be difficult to discern by direct inspection of crystal structures. In this context, it is particularly noteworthy that it is much easier to override the conservation rules at individual hot spots with the thermostable protein than with its mesostable homolog.

What makes the thermostable protein more mutationally tolerant? In the absence of a crystal structure for MjCM, it is difficult to provide a definitive answer to this question. However, given the similar helical content of our starting proteins and the strict conservation of the catalytic residues and the lengths of the intervening sequences,
differences in their respective native structures are unlikely to account for our observations. Instead, due to cooperative folding of the N- and C-terminal portions of the protein, sequence changes in the N-terminal region may simply be better accommodated when the C-terminal segment that was not mutated is more stable, as is likely to be the case for the thermostable starting protein. As for the protein as a whole, the enhanced stability of the MjCM C-terminal fragment presumably reflects an amalgam of factors, particularly better packing of hydrophobic residues and additional networks of hydrogen bonds and ionic interactions.\textsuperscript{116}

In conclusion, our results confirm the utility of using a thermostable protein as a starting point for protein design, even when large numbers of amino acid substitutions are made. Although the \textit{E. coli} CM tolerates extensive mutation surprisingly well, its thermostable counterpart affords a modest, but readily discerned advantage with respect to isolating functional clones. Our findings are particularly significant in light of the observation that high mutational loads have been shown to be beneficial for identifying improved or novel functions in protein libraries.\textsuperscript{112,117} Given the vastness of sequence space, such an advantage could be decisive by increasing the chances of finding variants of a particular protein with improved physical properties or extended chemical capabilities.
3. Enzyme Engineering by *in silico* Design and *in vivo* Selection with the Minimized Amino Acid Set IKEA

3.1. Introduction

Simplified proteins constructed from subsets of the 20 standard amino acids have been shown to fold into well-defined three-dimensional structures. These include molecules designed *de novo* as well as extensively mutagenized natural proteins. Such systems are useful for investigating folding mechanisms and for identifying the underlying determinants of protein structure. They can also provide insights into the evolutionary origins of modern proteins.

Our group has recently exploited powerful genetic selection methods to simplify the sequence of an enzyme, the AroQ class chorismate mutase from *Methanococcus jannaschii* (MjCM). All α-helices in the protein were replaced with randomized modules constructed from a minimal alphabet of four polar (DENK) and four apolar (FILM) residues. Functional variants were then selected by their ability to complement a CM-deficiency *E. coli* strain. This *in vivo* selection scheme allowed more than $10^7$ clones to be examined simultaneously, providing a quantitative assessment of the fraction of restricted sequences compatible with a catalytically active helical bundle structure. Only the highly conserved active site residues and the amino acids in the loops were held constant. To replace these amino acids with members of the 8 amino acid alphabet, the same *in vivo* selection system was employed to afford an active enzyme constructed from only 9 amino acids (the 8 amino acids from the alphabet plus arginine, needed as active site residue).

Although these selection experiments demonstrate the feasibility of creating a simplified enzyme from an easily encoded but otherwise arbitrarily chosen set of building blocks, many alternative alphabets could conceivably afford active catalysts. Experimental determination of the best — or least complex — set of building blocks for the AroQ fold from among the many possibilities would represent a dauntingly laborious undertaking, even with a powerful selection method in hand. However, computational methods, which enable rapid evaluation of enormous numbers of sequences, have the potential to guide the design of such experiments and aid the analysis of the subtle patterns of variation and covariation in functional sequences.
The program DESIGNER selects from among all possible sequences those likely to stabilize a given protein 3D structure, represented by the detailed atomic model of its backbone. Here we compare DESIGNER’s ability to predict amino acid sequences able to fold into the EcCM structure with the results of independent selection experiments which were carried out in parallel using libraries that were based on the 8 amino acid alphabet FILM-DENK and also with previously untested four amino acid alphabets. One four amino acid alphabet, IKEA, is tested computationally and experimentally in greater detail, because it covers most basic features of amino acids in proteins, including a large and a small hydrophobic residue (I and A) as well as an acid and a base (E and K). The best scoring sequence from a computational predicted IKEA variant is tested experimentally and its biochemical properties (secondary structure, stability and enzyme-activity) are compared to the wild-type EcCM.

The good agreement between theory and experiment suggests that this computational approach can greatly facilitate the search for viable reduced alphabets for constructing simplified scaffolds possessing truly protein-like properties.

3.2. Results and Discussion

H1 Library Design

Because the calculations are most reliable when applied to a backbone from a high-resolution 3D structure, the structurally characterized chorismate mutase from E. coli (EcCM) was employed for the in vivo and in silico selection experiments rather than its thermostable homolog from M. jannaschii. We targeted for randomization the 42 amino acid long H1 helix that spans this dimeric protein — the same helix investigated in Chapter 2 (Figure 4). Libraries were designed in analogy to the previously described EcCM and MjCM library modules. Implementation of the design entailed holding residues 1, 11, 21, 28 and 39 constant, either because they are part of the active site or are needed for construction purposes, and randomizing the rest of the helix according to the binary pattern of hydrophilic and hydrophobic residues shown in Figure 8. In general, mixtures of the two polar amino acids lysine (K) and glutamate (E) or the two apolar amino acids isoleucine (I) or alanine (A) were allowed at each variable position. This
severely restricted four-letter alphabet is closely related to the KEIAG alphabet that has been successfully used to encode 71% of a functional SH3 domain. This amino acid set was subsequently predicted computationally to be the optimal five-letter protein folding alphabet, but this prediction has also been disputed. At positions 15 and 18, which were previously identified in the MjCM and EcCM selection experiments as providing potentially important second sphere interactions with active site residues, all four amino acids from the minimal set plus aspartate (D) were permitted experimentally. The same library design used for the in vivo selection experiment was also used for the in silico prediction of sequence, called Designer23.

Figure 8. Design of library IKEA H1. Sequence of the N-terminal half with the long helix H1, aligned with the binary patterned randomized IKEA H1 module used in the in vivo selection experiment and the in silico selection of Designer23. Assigned polar residues (blue) were replaced by a 1:1 mixture of E and K, assigned apolar residues (red) were replaced with a 3:1 mixture of I and A, and at positions 15 and 18, E, K, I, A and D were offered in the experimental libraries. The catalytic residues Arg11, Arg28 and Lys39 as well as Ile21, which was needed for the assembly of the library, were held constant.

In vivo Selection Results

The randomized H1 module was constructed from two synthetic oligonucleotides, each about 100 nucleotides long. The polar amino acids were specified by the degenerate RAA codon (R = A or G) in a 1:1 ratio, whereas isoleucine and alanine were respectively encoded by the ATT or GCG codons, introduced into the sequence in a 3:1 ratio using the “split and mix” technique. Insertion of the H1 cassette into a plasmid in frame with the downstream portion of the EcCM wild-type sequence yielded the desired gene library.
The library was electroporated into the CM-deficient *E. coli* strain KA12/pKIMP-UAUC to give ca. $3 \times 10^6$ transformants. The correct construction of the library was confirmed by sequencing 32 randomly picked library members. The probability of finding lysine at polar sites and isoleucine at apolar sites was found to be $0.55 \pm 0.06$ and $0.78 \pm 0.04$ (mean $\pm 95\%$ confidence interval), respectively, in good agreement with the expected values of 0.50 and 0.75.

Selection for CM activity on minimal media lacking phenylalanine and tyrosine afforded approximately 1000 functional clones. Roughly 1 in 3000 genuine library members thus appears to encode an active CM. This complementation rate is approximately 10-fold higher than for replacement of the H1 helix in EcCM with a randomized module based on the eight amino acid alphabet FILM-DENK (Chapter 2), but similar to that observed in the analogous experiment with the thermostable MjCM template using the same eight amino acid set (Chapter 2).\textsuperscript{60}

The sequences of 21 functional enzymes were randomly picked and retransformed into the selection strain to verify their *in vivo* CM-activity. These 21 sequences were analyzed and the amino acid distributions at the randomized positions are summarized in Figure 9.

**Figure 9.** Analysis of active IKEA H1 variants. (A) The designed pattern and (B) the amino acid composition of functional clones versus sequence position. Asterisks denote significant deviations from a random distribution (as determined by $\chi^2$ analysis, $P < 0.01$).
In general, the H1 helix is very tolerant to simplification, underscoring the redundancy of information needed to assure proper folding of the helix, dimer assembly, and catalysis. Nevertheless, a quarter of the randomized positions show a statistically significant preference for one or the other choices offered. In some cases, the origin of these trends is obscure, for example at position 13 where a glutamate is preferred over lysine or at position 27 where the reverse is true. In others, the observed preferences can be rationalized according to well-established principles, such as minimization of exposed hydrophobic surface (position 2), enrichment of negatively charged residues at the N-terminus of a helix (position 5), or avoidance of steric clashes (position 36).

More interesting are the five sites (18, 22, 25, 29 and 35) that were subject to particularly strong selection pressure. Of these, the hydrophobic isoleucines at positions 22, 25 and 35 appear to contribute to efficient packing of the protein. Ile35, for instance, is one of the interdigitating residues in the H1-H1’ coiled-coil interface. The fact that it also directly contacts bound ligand in the crystal structure probably accounts for its strict conservation in functional clones. Residues 22 and 25, on the other hand, are involved in docking the H1 helices onto the templating H2 and H3 helices, respectively. They are more highly conserved than most of the residues at the H1-H1’ interface, presumably because assembly of a functional enzyme requires an optimal fit of the H1 helices against the invariable H2-H3 template. In contrast, suboptimal choices at individual positions in the H1-H1’ interface might be compensated combinatorially at other positions in the coiled-coil.

The hydrophilic residues at positions 18 and 29 also appear to be important for stabilizing the active form of the protein. As seen in earlier selection experiments with MjCM and EcCM, aspartate is strongly preferred at the former (and to a much lesser extent at position 15). In the crystal structure of EcCM, Asp18 forms a salt bridge with Arg28’, an essential active site residue, helping to position it for efficient catalysis. Glutamate is presumably much less effective for this purpose because of geometric constraints. The strong selection of a cationic lysine at position 29 can be rationalized by the favorable electrostatic interactions its side chain can make with Asp83 and with Asp69’, which likely stabilize the folded structure. This preference was only seen in experiments with EcCM, but not with its thermostable homolog MjCM template, which does not have the analogous interactions.
In silico Selection Results

In parallel to the in vivo experiments, in silico selection of amino acid sequences that are compatible with the given protein backbone structure of EcCM,85 were performed by Alfonso Jaramillo, Shoshana Wodak and Martin Karplus with the computer program DESIGNER.125 These calculations were done with binary patterned libraries of assigned hydrophilic and hydrophobic positions and with three different restricted sets of amino acids. First, the amino acid set FILM-DENK was used with the same binary pattern as was used for the in vivo selection experiment (Figure 6, EcCM library). Second, the amino acid set IKEA was used with a different binary pattern to match the pattern of the corresponding in vivo selection experiment. Residues 15D and 18D were fixed because of their functional role.60 Third, the amino acid set FGRD was used with the same binary pattern as the IKEA experiment. The sequence with the lowest calculated energy (the best scoring sequence) from the IKEA design (-316 kcal/mol) was significantly better than that from the FGRD design (-256 kcal/mol) and that from the FILM-DENK design (-253 kcal/mol).

The superiority of the IKEA design set in the in silico selection corresponds to the results of in the in vivo selection experiments. The IKEA design yielded roughly 1 functional sequence in 3,000 library members, whereas the FILM-DENK design yielded less frequent functional sequences (approximately 1 in 50,000). The counterintuitive result that the more restrictive IKEA design, with just 4 amino acids, performs better than the FILM-DENK design with 8 amino acids, is probably due to the exclusion of alanine in the FILM-DENK set. Analyzing the energies of every residue in the computed sequences showed that destabilizing positions in the FILM-DENK design contained an alanine in the IKEA design. Alanine's better performance in this helical structure is perhaps due to its high \( \alpha \)-helical propensity.126,127

Interestingly, the lowest energy of an IKEA sequence calculated with a fixed backbone (-316 kcal/mol; Table III) was below the average of the calculated energies of the in vivo selected IKEA sequences (-223 ± 28 kcal/mol (mean ± 95% confidence interval); Table IV).
Table III. Best scoring sequences from the IKEA library H1, selected *in silico*.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAEEKRIEIRDKIDDEEIIKKAERKKAIAEAAAKAE...</td>
<td>-315.6</td>
</tr>
<tr>
<td>MAEEKRIEIRDKIDDEEIIKKAERKKAIAEAAAKAE...</td>
<td>-315.2</td>
</tr>
<tr>
<td>MAEEKRIEIRDKIDDEEIIKKAERKKAIAEAAAKAE...</td>
<td>-315.2</td>
</tr>
<tr>
<td>MAEEKRIEIRDKIDDEEIIKKAERKKAIAEAAAKAE...</td>
<td>-314.8</td>
</tr>
</tbody>
</table>

*a* To the right of the sequences are their calculated energies (in kcal/mol). For this calculation (with the program DESIGNER) the same energy parameters were used and every sequence was mounted into the same backbone for all sequences in Table III, Table IV and Table V.

Table IV. Functional sequences from the IKEA library H1, selected *in vivo*.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAKKEKEEEIIEEREEAIKDDEEIIKKAERKKAIAEAAAKAE...</td>
<td>-221</td>
</tr>
<tr>
<td>MAEEKEKEEIIKIDKEEIIKAKRKEIIIKIKAIAKAIK...</td>
<td>-310</td>
</tr>
<tr>
<td>MAEEKEKEEIIKIDKEEIIKAKRKEIIIKIKAIAKAIK...</td>
<td>-187</td>
</tr>
<tr>
<td>MAEEKEKEEIIKIDKEEIIKAKRKEIIIKIKAIAKAIK...</td>
<td>-210</td>
</tr>
<tr>
<td>MIKKEKEEIIKIDKEEIIKAKRKEIIIKIKAIAKAIK...</td>
<td>-291</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-329</td>
</tr>
<tr>
<td>MIKKEKEEIIKIDKEEIIKIIAKRKEIIIKIKAIAKAIK...</td>
<td>-275</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-310</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-206</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-190</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-331</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-206</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-166</td>
</tr>
<tr>
<td>MIKKEKEEIIKIDKEEIIKIIAERKKAIAEAAAKAE...</td>
<td>-145</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-189</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-161</td>
</tr>
<tr>
<td>MIKKEKEEIIKIDKEEIIKIIAERKKAIAEAAAKAE...</td>
<td>-152</td>
</tr>
<tr>
<td>MAEEKEEKEEEIIKIAEIAEIIKIIAERKKAIAEAAAKAE...</td>
<td>-186</td>
</tr>
<tr>
<td>MIKKEKEEIIKIDKEEIIKIIAERKKAIAEAAAKAE...</td>
<td>-288</td>
</tr>
<tr>
<td>MIKKEKEEIIKIDKEEIIKIIAERKKAIAEAAAKAE...</td>
<td>not calculated</td>
</tr>
</tbody>
</table>

*a* See Footnote in Table III.

Table V. Randomly picked sequences from the library IKEA H1.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIKKEEKAERKKEIIKDDEEIIIKERKEIIKIIKIAIK...</td>
<td>- 77</td>
</tr>
<tr>
<td>MAEEKEKEIIEERKEIIKIDDEEIIKIIKIAIK...</td>
<td>-134</td>
</tr>
<tr>
<td>MAEEKEKEIIEERKEIIKIDDEEIIKIIKIAIK...</td>
<td>- 74</td>
</tr>
<tr>
<td>MIKKEEKEIIEERKEIIKIDDEEIIKIIKIAIK...</td>
<td>-220</td>
</tr>
</tbody>
</table>

*a* See Footnote in Table III.
To refine the search for the best scoring sequence with the program DESIGNER another calculation was performed in parallel with the *in vivo* experiments. The main difference to the other *in silico* selections was the use of a relaxed backbone. The same IKEA design was used as in the *in vivo* IKEA H1 library, including the randomization of position 15 and 18 (Figure 8). After 23 computational iterations, the sequence with the best energy score (-480 kcal/mol; MIKEKKIAEIRKKIEKIDEEIIKAAAERKKAAIEAAAKIE...) was obtained. It was dubbed "Designer23" and tested experimentally. A direct comparison of the calculated energy of Designer23 with the *in vivo* selected sequences is difficult, because this calculation does not take into account the energy of an unfolded or an alternative misfolded conformation, which might be different for the partially different sequences. Furthermore, Designer23 was designed to fit well in the given backbone, whereas the sequences from the *in vivo* experiment might have afforded much lower calculated energies if their backbone had been relaxed as well.

**Designer23 Characterization**

The computationally predicted sequence Designer23 was translated into a DNA sequence (see Appendix 7.1.7.). The gene encoding this protein and an appended C-terminal (His)$_6$-tag for efficient purification was expressed in a CM-deficient *E. coli* strain to prevent contamination by chromosomal CMs. After purification with affinity chromatography on Ni$^{2+}$-NTA agarose, the protein was characterized biochemically.

The secondary structure of Designer23 was investigated by CD spectroscopy (Figure 10). The CD spectrum shows the characteristic double minimum at 208 and 222 nm expected for a highly $\alpha$-helical protein and has the same shape as wild-type EcCM. The somewhat larger molar ellipticity compared to EcCM might reflect slight divergence in structure (a higher degree of $\alpha$-helicity in Designer23), but could also be due to differences in the determination of protein concentration (in both cases determined with BCA). To test the stability of Designer23, the change in ellipticity at 222 nm with temperature was measured (Figure 11).
Figure 10. Far-UV CD-spectrum of Designer23. Θ represents the molar ellipticity per residue.

Figure 11. Temperature-dependent unfolding (denaturation) curve of Designer23. The molar ellipticity Θ was measured at 222nm. T_m was calculated from the maximum of the first derivative with respect to temperature.75

The shape of the thermal denaturation curve suggests cooperative unfolding and an unfolding mechanism that is more complex than a simple two-state model (native dimer $\rightarrow$ 2 unfolded monomer). A clear transition is observed around 72°C.
(Figure 11), but an ill-defined transition might occur at lower temperatures, possibly due to partial unfolding of the protein. The apparent "melting temperature" $T_m$ of Designer23 (72°C) is significantly larger than the $T_m$ of EcCM (63°C),\textsuperscript{75} suggesting that the protein is more stable than its natural counterpart.

The catalytic efficiency of Designer23 ($k_{cat}/K_m = 1,400 \text{ M}^{-1}\text{s}^{-1}$) is lower than that of wild-type EcCM ($k_{cat}/K_m = 2.3 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$)\textsuperscript{128} by a factor of 160. For comparison, some of the selected FILM-DENK mutants were less active.\textsuperscript{60} Although only stability was modeled with DESIGNER, the in silico prediction yielded a very active enzyme. The large Michaelis constant of Designer23 ($K_m > 2500 \mu\text{M}$) confirms that the measured kinetics are not due to contamination with wild-type EcCM ($K_m = 300 \mu\text{M}$)\textsuperscript{128}. The activity, $\alpha$-helicity and stability of Designer23 establish its successful design.

**H2-H3 Libraries**

The successful simplification of the first helix in EcCM encouraged further efforts to replace the other two helices with the IKEA building blocks, as well. Akin to the design for the N-terminal part of EcCM (Figure 8), a binary pattern for the other half of the protein was designed (Figure 12).

![Figure 12](image)  
**Figure 12. Design of IKEA H2-H3 libraries.** The library modules are shown below the sequence of the wild-type EcCM. No gap is between this sequence and the N-terminal sequence shown in Figure 8. Lines above the sequences show the lengths of the H2 and H3 helices. Red and blue circles correspond to apolar and polar positions, where I + A and K + E were offered, respectively. Black circles indicate positions where I, K, E, A + D were offered. Capital letters indicate amino acids held constant by design. The IKEA H2-H3 library includes the entire sequence shown with the indicated positions randomized, whereas in the IKEA H3 library the H2 helix was kept constant and only the H3 segment shown on the right was randomized.
Figure 13. Strategies for enzyme-simplification. In black the simplified helices, in white the wild-type sequences of EcCM. Four different strategies are possible: A) direct, B) stepwise, C) with more steps, D) in parallel. The first step in the strategies B, C and D is the same and involves simplification of the long H1 helix. Strategy D was previously used successfully with the larger FILM-DENK amino acid set and a thermostable homolog of EcCM.60

To simplify the entire enzyme, different strategies were elaborated (Figure 13). In principle, all the helices of EcCM could be randomized simultaneously (A in Figure 13), but the required library sizes are larger than the experimentally accessible library sizes.60 Consequently, stepwise strategies were developed (Figure 13). The first strategy to be employed (B in Figure 13) started from simplified H1 variants. Fourteen active IKEA H1 variants were simultaneously crossed with the IKEA H2-H3 library (Figure 12) to yield ca. \(1.6 \times 10^7\) library members after transformation. However, no active library member was found after selection. Consequently, a strategy with smaller steps (C in
Figure 13) was employed. The 14 active IKEA H1 variants were crossed with the IKEA H3 library (keeping the H2 wild-type sequence). This experiment yielded after transformation ca. $1.8 \times 10^7$ library members, but none of them was active either. The next strategy was to simplify the second half of the protein independently of the first, with the expectation that if active variants should be found, we would cross active IKEA H1 with IKEA H2-H3 variants in a subsequent experiment. Transformation of the IKEA H2-H3 library (keeping the H1 wild-type sequence) yielded ca. $1.2 \times 10^7$ library members, but again no functional sequences could be isolated. The correct construction of the libraries was confirmed by sequencing randomly picked library members. In a final experiment, the IKEA H3 library (keeping the H1 and H2 wild-type sequence) was transformed into the selection strain to give ca. $1.7 \times 10^7$ library members. Again no functional variants were selected, although an independent selection experiment (Chapter 4), which was done in parallel, yielded active CM enzymes.

A possible reason for the inability to obtain active IKEA H2-H3 or H3 variants by this approach could be the Q88E mutation that we introduced. The reason for including this conservative mutation from glutamine to glutamate, was that Q is not part of the IKEA alphabet, whereas E is. Although other active site residues are strictly conserved in wild-type CMs, Q and E are found at position 88 in an alignment of natural CMs. Furthermore, the active site position 88 could be mutated (without loss of *in vivo* activity) from the wild-type Q to E, N and D (E gave the fittest mutants), in a comparable simplification project with the homologous, but thermostable MjCM. These observations supported the design decision. However, subsequently done selection experiments (Chapter 4), where position 88 was randomized to E, Q, N, D, R, K, H, S and G in the same EcCM sequence, only yielded Q88 in all 110 active variants.

The Q88E hypothesis could be easily tested, if oligonucleotides synthesized with the "split and mix" technique were commercially available. A newly developed DNA synthesis method, called SlonoMax, which has been recently commercialized by Sloning BioTechnology GmbH (Puchheim, Germany), has the potential to deliver the oligonucleotides needed for IKEA libraries (Chapter 1.2.).

Another possible reason for the inability to obtain active IKEA H2-H3 and H3 variants could be that large, aromatic amino acids are missing in the IKEA set. In a comparable experiment with the homologous MjCM and the FILM-DENK set,
phenylalanine was highly enriched at position 77 in the helix H3, although at other positions a strong bias against phenylalanine was observed. Both wild-type enzymes, EcCM and MjCM contain phenylalanine at position 77. The F77 hypothesis could be easily tested by conserving the wild-type residue at position 77 instead of randomizing it. Both effects, Q88E and F77, and other effects could cumulatively decrease the chances of finding active variants.

**Perspective**

Numerous proteins have been mutagenized with the goal of replacing large portions of sequence with subsets of the 20 canonical amino acids. These include the SH3-domain, orotate phosphoribosyl transferase and chorismate mutase. These studies show that retention of function is possible with reduced amino acid sets. The results are consistent with the hypothesis that protein evolution started with a simpler amino acid set than the one used today.

Our results extend this work. We have been able to replace 40% of chorismate mutase with one of the 4 amino acids I, K, E or A. This set contains a large and a small aliphatic amino acid and a positively and a negatively charged residue. Furthermore, this set is a subset of the set IKEAG that was used to replace 71% of the SH3 domain.

As in the earlier study, we find, however, it was not possible to simplify the entire protein using such a minimal set. The precise reasons are unknown. As noted above, in the case of EcCM, mutations in the third helix, especially at positions 77 and 88, may have been critical. It is also clear that using reduced sets of amino acids can be problematic. Dispensing with a large number of amino acid types significantly reduces the diversity of favorable internal packing interactions. For example, the previously described 9 amino acid chorismate mutase, while active, exhibits many of the properties of a molten globule, which is substantial less stable and more wobbly than its counterpart, the highly evolved EcCM. We can expect that such problems will be more severe with a 4 amino acid set. In that case, the search for viable fully randomized sequences could be extremely difficult — even with our powerful selection system.
Given the excellent agreement between experiment and computation in the case of the H1 library, the obvious next step would be to extend the computational approach to identify functional H2 and H3 IKEA sequences. Although this experiment could not be carried out during this thesis due to time limitations, fully simplified IKEA sequences could be predicted computationally and tested experimentally the same way we simplified the H1 segment.

Of course, it is possible that a fully simplified and active chorismate mutase based on the IKEA alphabet cannot exist. Here, computation may help as well to evaluate alternative reduced alphabets. Computational methods are certainly faster than experimental methods and can consider many more alternative designs. We hope that the combination of experiment and computation will ultimately lead to a wide range of functional sequences with tailored properties.
4. Plasticity of an Enzyme Active Site

4.1. Introduction

Creating enzymes that catalyze unprecedented reactions is a major challenge, which would have numerous applications. Instead of de novo design, existing protein structures can be used as scaffolds for new active sites. To construct active sites, a detailed understanding of sequence-structure-activity relationship is needed — and in reverse, successful design is a good indication that the underlying principles are sufficiently understood.

The current study addresses the questions: (1) How "plastic" are modern, highly evolved enzyme active sites? (2) Do alternative active site architectures exist, that bind the transition state analog in a different orientation? (3) How frequent are such alternatives in libraries of randomized active site residues? To answer these questions, we have chosen EcCM as a model system, because a powerful survival-selection system is available.93

To investigate the architectures of active sites, enzyme cocrystallization with a transition state analog, alignment of homologous sequences, site-directed mutagenesis and computational modeling have been used.45,75,85,130,131 Such studies on EcCM have revealed that its active site is completely engulfed by the protein.85 Three arginines and one lysine dominate the positively charged active site, and are needed to match the negative charges of the substrate and the polarized transition state.85 Stabilization of the transition state, rather than ground state conformations, has been suggested to be the main reason for the catalytic efficiency of CM.45,47 Unambiguously, the catalytic mechanism of CM involves neither a covalent bond between enzyme and an intermediate, nor general acid or base catalysis.130 Cofactors are not involved in this simple one substrate to one product reaction either. These properties of CM make it an ideal model system for studying enzyme catalysis. Furthermore, EcCM's small gene size, encoding only 97 amino acids, and the availability of a genetic selection system for CM-activity, make it a good choice for our study.

To answer the questions posed above, we simultaneously mutated all active site residues in EcCM and selected for CM activity. This approach provided an internal positive control (at least one wild-type-like sequence should be found). Compared to
conventional site-directed mutagenesis and saturation mutagenesis of single residues, our approach also allows identification of synergistic and compensatory effects, determining in one experiment how many variants of the wild-type enzyme are active and exploring sequence space for alternative active site architectures (Figure 14).

![Figure 14. Metaphoric representation of sequence space as a desert landscape. (Monument Valley, USA) Most sequences are inactive (represented by the desert floor). Around one highly active sequence, some variants are also active (represented by the peak in the middle). Where are the completely different sequences with activity (represented by the second peak)? How frequent are they? How do we get there?](image)

### 4.2. Results and Discussion

**Library Design and Construction**

The active site residues of EcCM were targeted for randomization in two libraries. In the first, called VII, all 7 polar active site residues were randomized. All side-chains were chosen that make direct contact with the functional groups of the transition state analog by inspecting the crystal structure (Figure 15). At each position the following amino acids were offered: Arg, Lys, His, Glu, Gln, Asp, Asn, Ser and Gly in a 3:1:1:1:1:1:1:2 ratio. This subset of all 20 possible amino acids was chosen, because it can be easily encoded with a degenerate codon (VRM) and it contains all polar amino
acids except threonine and tyrosine, which are not part of the active site of wild-type EcCM. In the second library, called XII, all 12 active site residues (the 7 polar plus 5 apolar side-chains that contact the transition state analog) were randomized. The amino acids Arg, Lys, His, Gln, Asn, Ser, Ile and Leu were allowed at each position in a 3:1:1:1:1:1:1:2:2 ratio. This amino acid subset is related to the subset of library VII, but contains two hydrophobic residues (Ile and Leu). Focused subsets were used to increase the chances of finding functional sequences. The libraries were enriched with Arg codons (the active site of wild-type EcCM contains three arginines) and free of unwanted codons, including stop codons and codons for Pro (a helix breaker), Cys (which may form disulfide bridges) and other amino acids that are not present in the active sites of natural CMs.75

Figure 15. Interactions of active site residues from wild-type EcCM with a transition state analog. Seven hydrophilic residues (blue) and five hydrophobic residues interact with the transition state analog (green). The side-chains of these active site residues are in direct contact with the transition state analog and are within 6.5 Å of its first carbon.
**Figure 16. One-pot synthesis of library XII.** The top row shows the EcCM sequence. Asterisks indicate active site residues, which were randomized in library XII. Double stranded DNA used as template is shown in green; primers and products in blue. The 3’ ends of DNA are denoted with arrow heads.

To construct the designed libraries, a one-pot synthesis was developed (Figure 16) that makes purification of intermediates unnecessary and thereby minimizes hands-on effort and resources. Because the randomized positions are scattered over the whole coding sequence, and because an oligonucleotide corresponding to this sequence is too long for current synthetic methodology, the libraries were constructed from two primer pairs. More than two primer pairs can be used to extend this approach to longer libraries. Each primer pair leads to a PCR product, which is an intermediate for the final PCR. Each intermediate is a double stranded DNA that contains one strand that can act as a primer (called a megaprimer) for the final PCR. To ensure incorporation of the inner primers and to avoid contamination with wild-type template, two different templates were used for the two primer pairs. Furthermore, silent mutations were introduced at the 5’ ends of the inner primers to avoid annealing of the 3’ ends of the megaprimers with the templates. Thus, the megaprimer can only prime themselves, which leads to the desired product. To favor synthesis of the megaprimer and consequently the final product over the unproductive second strand of the intermediates, the outer primers were added in excess over the inner primers.
**Library Characterization**

The experimental library sizes obtained after electroporation into a CM deficient strain are on the same order of magnitude as the theoretical library size of library VII with its 7 randomized polar active site residues (Table VI). In contrast, the 12 randomized active site residues of library XII result in a theoretical size that is beyond experimentally accessible library sizes (Table VI). For library VII, the complementation frequency (the number of selectants divided by the experimental library size) is about 1 in 100,000 and was reproducible in two independent experiments (Table VI).

**Table VI. Library statistics.**

<table>
<thead>
<tr>
<th>Library:</th>
<th>VII</th>
<th>XII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical library sizes:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>at the amino acid level:</td>
<td>$9^7 = 4.8 \times 10^6$</td>
<td>$8^{12} = 6.8 \times 10^{10}$</td>
</tr>
<tr>
<td>at the codon level:</td>
<td>$12^7 = 3.6 \times 10^7$</td>
<td>$12^{12} = 8.9 \times 10^{12}$</td>
</tr>
<tr>
<td>Experimental library sizes:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment A:</td>
<td>$1.4 \times 10^7$</td>
<td>$1.7 \times 10^7$</td>
</tr>
<tr>
<td>Experiment B:</td>
<td>$4.1 \times 10^7$</td>
<td></td>
</tr>
<tr>
<td>Number of selectants:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment A:</td>
<td>144</td>
<td>1</td>
</tr>
<tr>
<td>Experiment B:</td>
<td>378</td>
<td></td>
</tr>
<tr>
<td>Complementation frequency:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment A:</td>
<td>1 in 97,000</td>
<td></td>
</tr>
<tr>
<td>Experiment B:</td>
<td>1 in 108,000</td>
<td></td>
</tr>
</tbody>
</table>

**Table VII** shows that more than 50% of the randomly picked library members (non-selectants) were as designed, and that the percentages correspond for the two libraries synthesized with two different sets of oligonucleotides. Most of the other non-selectants contained deletions. Other artifacts (single base insertions and substitutions or plasmids that still contained the stuffer fragment or lacked an insert) were found rarely. The point
mutations probably arose during chemical synthesis of the oligonucleotides, and the other rare artifacts during plasmid digestion and ligation. Surprisingly, 25/110 (23%) of the functional sequences had mutations in the region of the strong, constitutive trc promoter, compared to only 1/30 (3%) of non-selectants. The fraction of correct non-selectants is very similar for both libraries. Although it would be desirable to have a higher proportion of correct library members, the maximally theoretically possible increase from 50% to 100% would only double the library size. In contrast, improvement in transformation efficiency tripled the library size (Table VI). Optimization of electroporation efficiency is easier than improvement of the already highly optimized, commercial synthesis of oligonucleotides.

Table VII. Sequence analysis.

<table>
<thead>
<tr>
<th>Library:</th>
<th>VII</th>
<th>XII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-selectants:</td>
<td>53% (16/30) correct</td>
<td>55% (6/11) correct</td>
</tr>
<tr>
<td></td>
<td>47% (14/30) artifacts (mainly deletions)</td>
<td>45% (5/11) artifacts (mainly deletions)</td>
</tr>
<tr>
<td>Selectants:</td>
<td>79% (110/140) positives</td>
<td>1 positive</td>
</tr>
<tr>
<td></td>
<td>21% (30/140) artifacts (mainly deletions, mixtures of two plasmids)</td>
<td></td>
</tr>
</tbody>
</table>

The one active sequence selected from library XII, where all 12 active site residues were randomized, contains just two mutations relative to the EcCM wild-type sequence (Arg11Ile and Ile14Lys). The codon usage of this mutant reveals that the second half of the coding sequence probably originated from the parent gene and only the first half was randomized as designed. All 170 sequenced clones from library VII did not reveal a comparable case. All sequenced non-selectants from library XII were fully randomized. Because the experimental size of library XII is $1.7 \times 10^7$, it can be concluded that the frequency of functional clones in this population is less than 1 in $10^7$ (Table VI).

In contrast to the results with library XII, a significant number of functional
sequences was obtained from library VII. Some of the selected clones from library VII contained mixtures of two plasmids, which were easily detected in the chromatograms obtained from the sequencer (Figure 17). Such mixtures could arise from concurrent transformation of one cell with two plasmids, which would be favored by the high DNA concentrations used here for electroporation or by transformation with heteroduplex DNA.64

Figure 17. Sequencer chromatograms of two different samples illustrating plasmid mixtures. Each sample was isolated from one selected clone and contains a mixture of two plasmids. The main and the minor signals are out of synchronization at the right side of the chromatograms as a consequence of a single-base deletion in the main sequence. The minor blue peak at count 230 occurs at the first randomized position (amino acid 11). At this position, the main plasmid in A has the codon GGA, which encodes glycine, while the minor plasmid contains CGA for arginine. In B, amino acid 11 is a glutamate (GAA) in the main sequence, and Arg (CGA) in the minor. An arginine at position 11 is crucial for enzyme activity.130 The minor plasmids also contain the expected codons for an active enzyme at the other randomized positions. The main plasmids encode truncated proteins (which are only 21 amino acids long and are unlikely to have CM-activity) due to the frameshift mutations.
The codon usage in library members can be used to monitor the selection process. Prior to selection, codon usage is randomly distributed within the design as seen in the 16 correct length non-selectants from library VII, analyzed in Table VIII.

Table VIII. Codon usage for 16 correct length non-selectants, all 110 sequenced positives from library VII, and the EcCM wild-type gene. The fact that no codons except for Arg were found at positions 11 and 28 after selection is highlighted in yellow.

<table>
<thead>
<tr>
<th>Sequence position:</th>
<th>11</th>
<th>28</th>
<th>39</th>
<th>51</th>
<th>52</th>
<th>84</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 non-selectants:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Asn</td>
<td>3</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Ser</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td></td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Asp</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gly</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>All 110 positives:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>109</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td></td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td></td>
<td>1</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td></td>
<td>6</td>
<td>12</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td></td>
<td></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>36</td>
<td>16</td>
<td>1</td>
<td>34</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>63</td>
<td>50</td>
<td>41</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGC</td>
<td>10</td>
<td>44</td>
<td>26</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGG</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wild-type:</td>
<td>CGA</td>
<td>CGG</td>
<td>AAA</td>
<td>CGT</td>
<td>GAA</td>
<td>TCC</td>
<td>CAG</td>
</tr>
</tbody>
</table>

46
After selection for functional sequences, only arginine codons were selected at position 11 and 28 (Table VIII). However, all three designed arginine codons were found and, in rare cases (with less than 1% frequency), even arginine codons excluded by design were found. The difference in codon usage for library members with and without selection pressure confirmed that genuine selection had occurred. The distinct codon usage in selected sequences that encode the same protein demonstrates the efficient incorporation of the degenerate inner primers. Furthermore, silent mutations designed into all primers establish that every selected sequence originated from the library oligonucleotides and not an external source. No contamination with wild-type genes was observed. However, two sequences were selected that are like the EcCM wild-type sequence, but use different codons than the PCR-template. In addition to the 2% (2/110) wild-type-like sequences, 85% (93/110) of the variants that were found in library VII have one mutation and 14% (15/110) have two mutations at the amino acid level.

**Analysis of Active Sequences**

Table IX summarizes the analysis of all 93 sequenced positives that had one active site mutation (wild-type and functional sequences with more than one mutation are not included). As expected, some active site residues (11, 28 and 88) were completely resistant to point mutation. At other sites, conservative changes (e.g. between positively charged residues at positions 39 and 51) were tolerated in rare cases (with 1% frequency). A third category of sites accepted change readily (positions 84 and especially 52), even though these residues are in direct contact with the transition state analog in the crystal structure. In natural CMs of the AroQ-class Arg11, Arg28, Lys39, Arg51 and, interestingly, also Glu52 are strictly conserved, whereas Ser84 is not. At position 88, both Gln and Glu are found in natural CMs.
Table IX. Analysis of 93 positives with one mutation in the active site.

<table>
<thead>
<tr>
<th>Sequence position:</th>
<th>11</th>
<th>28</th>
<th>39</th>
<th>51</th>
<th>52</th>
<th>84</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>99%</td>
<td>48%</td>
<td>95%</td>
<td>100%</td>
</tr>
<tr>
<td>Positives:</td>
<td>Arg</td>
<td>Arg</td>
<td>Lys</td>
<td>Arg</td>
<td>Gly</td>
<td>Ser</td>
<td>Gln</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>1%</td>
<td>26%</td>
<td>1%</td>
<td>10%</td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Lys</td>
<td>Arg</td>
<td>Lys</td>
<td>Gly</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8%</td>
<td>4%</td>
<td>2%</td>
<td>2%</td>
<td>1%</td>
<td>1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td>Gln</td>
<td>Asn</td>
<td>Asn</td>
<td>Asp</td>
<td>Ser</td>
<td></td>
</tr>
<tr>
<td>Wild-type:</td>
<td>Arg</td>
<td>Arg</td>
<td>Lys</td>
<td>Arg</td>
<td>Glu</td>
<td>Ser</td>
<td>Gln</td>
</tr>
</tbody>
</table>

In EcCM, Glu52 provides a hydrogen bond to the 4-hydroxyl group of the transition state analog (Figure 15). Although Glu52 is strictly conserved in natural CMs, it is evidently not necessary for in vivo CM activity (Table IX). In this context, EcCM can be compared to isochorismate pyruvate lyase from Pseudomonas aeruginosa (PaIPL). These enzymes are structural homologs that share a similar active site architecture and mechanism, but catalyze different reactions. Despite the low sequence identity between EcCM and PaIPL of less than 20%, they share five polar active site residues (Arg11, Arg28, Lys39, Arg51, Gln88; EcCM numbering), but not Glu52 or Ser84, the most variable residues in our selection experiment (Table IX). Glu52 and Ser84 were mutated on their own (Table IX) and also simultaneously (Table XI), without loss of CM function. Glu52 is replaced by Val in PaIPL, whose substrate isochorismate lacks the 4-hydroxyl group of chorismate (Figure 18), whereas Ser84 is replaced by Tyr, whose bulky side chain points away from the active site. Although the residue at
position 84 has no obvious functional role, Glu52 may be conserved in natural CMs to prevent wasteful side reactions, like the conversion of chorismate to p-hydroxybenzoate or the transformation of isochorismate to salicylate.

![Chorismate and isochorismate](image)

**Figure 18.** Chorismate and isochorismate, the substrates for chorismate mutase (CM) and isochorismate pyruvate lyase (IPL), respectively.

Because isochorismate pyruvate lyase from *P. aeruginosa* is a promiscuous enzyme with modest CM activity, it was speculated that PaIPL might have evolved from a CM.132 Wild-type EcCM has no measurable IPL activity,134 but the mutational tolerance of EcCM at position 52 and 84 (Table IX and Table XI) might have provided the necessary flexibility to convert EcCM into a promiscuous CM-IPL. Evolution of a new function is often driven by mutations that do not abolish the original function but, at the same time, enable promiscuous functions, thereby allowing evolution of new activity prior to gene duplication.135,136
Table X  All previously published mutants of EcCM that are theoretically accessible with library VII.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>$K_m$  (μM)</th>
<th>$k_{\text{cat}}$ (s$^{-1}$)</th>
<th>$k_{\text{cat}} / K_m$ (M$^{-1}$ s$^{-1}$)</th>
<th>Selected from library?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type a</td>
<td>300</td>
<td>72 d</td>
<td>240 000 d</td>
<td>yes</td>
</tr>
<tr>
<td>Glu52Gln a</td>
<td>1100</td>
<td>24</td>
<td>23 000</td>
<td>yes</td>
</tr>
<tr>
<td>Glu52Asp a</td>
<td>1400</td>
<td>3</td>
<td>2 200</td>
<td>yes</td>
</tr>
<tr>
<td>Arg11Lys a</td>
<td>&gt;2000</td>
<td>230</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Arg28Lys a</td>
<td>&gt;2000</td>
<td>230</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Gln88Glu a</td>
<td>&gt;2000</td>
<td>360 c, d</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Gln88Lys a</td>
<td>&gt;2000</td>
<td>12</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Lys39Gln a</td>
<td>&gt;2000</td>
<td>7</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Lys39Arg a</td>
<td>&gt;2000</td>
<td>2 d</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>Wild-type b</td>
<td>300</td>
<td>41 d</td>
<td>140 000 d</td>
<td>yes</td>
</tr>
<tr>
<td>Lys39Arg b</td>
<td>590</td>
<td>0.12</td>
<td>210 d</td>
<td>yes</td>
</tr>
<tr>
<td>Lys39Asn b</td>
<td>450</td>
<td>0.05</td>
<td>110</td>
<td>no</td>
</tr>
<tr>
<td>Lys39Gln b</td>
<td>1200</td>
<td>0.01</td>
<td>9</td>
<td>no</td>
</tr>
<tr>
<td>Gln88Glu b</td>
<td>&gt;1600</td>
<td>310 c, d</td>
<td></td>
<td>no</td>
</tr>
</tbody>
</table>

a  measured at 30°C in 50 mM phosphate buffer pH = 7.5

b  measured at 37°C in 50 mM Tris buffer pH = 7.8

c  strongly pH dependent

d  differences in the published values might not be only due to the different assay conditions, but also due to different methods in determining protein concentrations

Among the positives from our selection experiments, all previously published active site mutants were found if their activity was above the threshold of the selection system (Table X). This threshold depends not only on catalytic efficiency ($k_{\text{cat}} / K_m$), but also on 50
protein concentration in the cell, which in turn depends on expression efficiency and on protein stability (or half-life). Assuming similar protein concentrations and an equal distribution of each sequence in the library, and correcting for library quality and the number of sequenced clones, we estimate that more than two thirds of all possible 7 amino acid variants with more than ca. 0.1% of wild-type CM-activity have been recovered from library VII. To achieve a 95% chance of complete coverage of library VII, an experimental library size would be required that is more than 18 times larger than the theoretical library size.

Besides the single mutants described above (Table IX and Table X), active variants were also found with two mutations at the active site (Table XI and Table XII). Positions 52 and 84, which turned out to be the most receptive positions for single mutation (Table IX), were also receptive to simultaneous mutation without loss of function (Table XI). Another class of double mutants involves the adjacent residues 51 and 52 (Table XII). That a positive charge, provided by residue 51, is needed for activity is apparent from the analysis of single mutants (Table IX), but this positive charge could be provided from the adjacent position 52 in the double mutant Arg51Gln/Glu52Arg. This double mutant was found independently five times, each time with different codon usage. Another interesting double mutant is Arg11Ile/Ile14Lys, which is the one functional sequence from library XII. Here again, the need for a positive charge provided by residue 11 is obvious from Table IX. In the Arg11Ile/Ile14Lys mutant, residues 11 and 14 seem to have switched function. This kind of covariational mutation is only accessible with simultaneously randomized positions, and not with conventional site-directed mutagenesis or saturation mutagenesis.
Table XI. Mutations at positions 52 and 84 that occurred simultaneously and independently without loss of function.

<table>
<thead>
<tr>
<th>Double mutants</th>
<th>Single mutants</th>
<th>Selected from library?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu52Gln / Ser84Gly a</td>
<td></td>
<td>yes, 7 times</td>
</tr>
<tr>
<td>Glu52Gln / Ser84Asn b</td>
<td></td>
<td>yes, 1 time</td>
</tr>
<tr>
<td>Glu52Gln a,b</td>
<td></td>
<td>yes, 4 times</td>
</tr>
<tr>
<td>Ser84Gly a</td>
<td></td>
<td>yes, 2 times</td>
</tr>
<tr>
<td>Ser84Asn b</td>
<td></td>
<td>yes, 3 times</td>
</tr>
</tbody>
</table>

a,b Double mutants and their corresponding single mutants.

Table XII. Mutations at positions 51 and 52 that occurred simultaneously, but the corresponding mutations at position 51 did not occur independently.

<table>
<thead>
<tr>
<th>Double mutants</th>
<th>Single mutants</th>
<th>Selected from library?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg51Gln / Glu52Arg a</td>
<td></td>
<td>yes, 5 times</td>
</tr>
<tr>
<td>Arg51Gln / Glu52Asn b</td>
<td></td>
<td>yes, 1 time</td>
</tr>
<tr>
<td>Arg51His / Glu52Arg c</td>
<td></td>
<td>yes, 1 time</td>
</tr>
<tr>
<td>Arg51Gln a,b</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Arg51His c</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Glu52Arg a,c</td>
<td></td>
<td>yes, 24 times</td>
</tr>
<tr>
<td>Glu52Asn b</td>
<td></td>
<td>yes, 2 times</td>
</tr>
</tbody>
</table>

a,b,c Double mutants and their corresponding single mutants.
Characterization of Selected Clones

Unexpected sequences, which have not been seen previously, were characterized in more detail. First, positives were retransformed into the selection strain to verify the \textit{in vivo} activities (Table XIII). After retransformation, the sequences of all clones were confirmed.

Table XIII. \textit{In vivo} CM activity of retransformed clones.

<table>
<thead>
<tr>
<th></th>
<th>Growth on minimal medium with Tyr</th>
<th>Growth on minimal medium without Tyr</th>
<th>\textit{In vivo} CM activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative controls (^a)</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Positive control (^a)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Arg51Lys</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Arg11Ile / Ile14Lys</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Glu52Gly</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Glu52Arg</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Arg51Gln / Glu52Arg</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Arg51Gln / Glu52Asn</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Arg51His / Glu52Arg</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

\(^a\) The controls contained the same vector sequence as the other samples and differed only in the coding sequence, which are EcCM mutants from — and tested in — previous experiments (described in the previous chapter).

Once the phenotype was confirmed, individual proteins were produced and purified. To enable efficient production of enzymes, the coding sequence from selected library plasmids was subcloned into an expression vector. The vector constructs were transformed into CM-deficient strain KA12, sequenced to confirm their correct construction, and retransformed into the CM-deficient expression strain KA13 for protein
production. Proteins were produced with a C-terminal (His)$_6$-tag and purified on an affinity column with immobilized nickel. Enzyme kinetics were measured by observing the consumption of chorismate photometrically (Table XIV). All measured enzymes had $K_m$ values larger than 2500 $\mu$M, except wild-type EcCM, which has a $K_m$ of ca. 300 $\mu$M as previously reported,$^{128,130,131}$ showing that the measured activities are not due to contamination with wild-type EcCM. All measured proteins were active in vitro, as expected from the in vivo experiments. The calculated $k_{cat} / K_m$ values are summarized in Table XIV.

### Table XIV. In vitro CM activity of purified variants.

<table>
<thead>
<tr>
<th>Variant</th>
<th>$k_{cat} / K_m$ (M$^{-1}$ s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type$^{128}$</td>
<td>2.3 $\times$ 10$^5$</td>
</tr>
<tr>
<td>Arg51Lys</td>
<td>1570</td>
</tr>
<tr>
<td>Arg11Ile / Ile14Lys</td>
<td>230</td>
</tr>
<tr>
<td>Glu52Arg</td>
<td>240</td>
</tr>
<tr>
<td>Arg51Gln / Glu52Arg</td>
<td>58</td>
</tr>
<tr>
<td>Arg51His / Glu52Arg</td>
<td>57</td>
</tr>
</tbody>
</table>

**Conclusions**

In total, 16 different protein sequences from library VII, where all 7 polar active site residues of EcCM were randomized, were found with CM-activities of biological significance. This variability of the EcCM active site is mostly due to two active site residues (52 and 84). The other residues and the active site as a whole were more resistant to change, which restricts the "plasticity" of this active site. Apparently, it is not possible to radically redesign this binding pocket without substantial losses in CM activity (> 10$^3$-fold).
The reason why variants of the wild-type EcCM — and not alternative active sites architectures — were found in library VII, can be attributed to a memory effect. This effect might be due to the trajectories of the randomized side-chains, which are fixed by the backbone and thus specified by this scaffold. Also the 5 unchanged hydrophobic residues that provide a template into which residues from the library must slot, have limited the extent of change possible.

To circumvent such a slot effect, the 5 hydrophobic active site residues were randomized in addition to the 7 polar active site residues, affording library XII. Through serendipity, one active EcCM variant (Arg11Ile/Ile14Lys) was selected from library XII. However, its DNA sequence (but not its amino acid sequence) indicates that it is outside the library XII design. Like a positive control experiment, this successful selection shows that the selection assay worked properly — also with library XII. Because no other functional sequence (not even wild-type) was found in 17 million library XII members, it can be concluded that new active site architectures are extremely rare (< 1 in 10⁷ with this library), if they exist at all with this scaffold.

The large dynamic range of the applied selection system (illustrated in Table XIV) would have allowed the identification of alternative active site architectures, even if they would confer only marginal activity (~ 10³-times lower than wild-type). Furthermore, this survival-selection system enabled testing larger libraries than with high throughput screening. Apparently, genetic selection or high throughput screening from randomized libraries are unlikely to generate new active sites, even for reactions as simple as CM, even when the starting structure is ideally suited for the reaction, and even when the library is biased to favorable amino acid compositions.

How can we create new active sites for new reactions? Additional information — which can be encoded in primary sequences — will evidently be essential. This information can be implemented through computationally guided library design or computing lead sequences that are improved by directed evolution.¹⁰ The advantage of computation is that much larger libraries can be assessed than what would be possible experimentally (Chapter 3). The advantage of library approaches can then be utilized to fine-tune the subtle requirements of an efficient enzyme and to increase the chance of finding functional sequences compared to testing a handful predictions.
5. Outlook

In the first part of this thesis (Chapters 2 and 3) the relationship between sequence and function was explored by systematic variation of long segments of the enzyme chorismate mutase. Specifically, modules of randomized sequence were used to replace secondary structure elements. In the complementary experiments described in Chapter 4, the active site residues of the enzyme chorismate mutase were randomized. Functional variants within these libraries were identified by their ability to rescue a chorismate mutase-deficient E. coli strain under auxotrophic conditions. The first key finding of this thesis is that a thermostable protein is superior to its mesostable homolog with respect to its use as a starting structure for enzyme engineering, even under high mutational load (Chapter 2). Second, the successful prediction of a functional sequence, using the 4 amino acid alphabet IKEA shows that computer models can aid in the quest for the minimal set of amino acids (Chapter 3). Third, alternative active sites appear so rarely in sequence space that they are only found if additional information is included in the library design (Chapter 4). This information can be in the form of binary patterning for example (Chapter 1.2.) to favor the formation of folded structures. Alternatively, other structural biases and mechanistic insights can be used to increase the chance of finding functional sequences.

The library-based approaches and the use of reduced amino acid sets that are employed in this thesis build on earlier studies. The impressive developments in this research area are illustrated by the following time-line. In 1988, a library-based method was first used to probe protein structure-function relationships. In 1993, binary patterning was successfully applied to design a four-helix bundle protein that folds properly. In 1997, another four-helix bundle protein was constructed from a 7 amino acid alphabet that had the intended structure. Also in 1997 it was shown that 71% of a small β-sheet protein could be replaced by a 5 amino acid alphabet with a library-based approach, without loss of its binding capacity. In 2001, library-based approaches and function-selection systems were employed to mutate triosephosphate isomerase and chorismate mutase with a 7 and 8, respectively amino acid set. In 2002, 88% of the enzyme orotate phosphoribosyl transferase was replaced by a 9 amino acid alphabet with a library-based approach, while retaining its activity. In 2005, the whole sequence of
an α-helical chorismate mutase was replaced by a 9 amino acid alphabet to afford an active enzyme that functions in living cells.\textsuperscript{55} As this time-line shows, it has been possible to design more and more sophisticated protein properties, starting from folded structures, progressing to binders and, most recently, creating active catalysts. In this thesis, enzyme activity was also chosen as the most sensitive readout of successful design. Although we have known for about ten years that not all 20 amino acids are necessary to construct functional proteins, we still do not know which subset is the optimal set for a given task, or how few amino acids are really needed, or which amino acids constituted the first proteins on earth. Our study has tested the most severely reduced amino acid set so far for constructing active enzymes. As shown in Chapter 3, these efforts will certainly benefit from computation. Such studies can, in turn, reduce computational complexity and thus facilitate computer-guided enzyme engineering. They may also help understand early protein evolution, which is believed to have started with a simpler set of amino acids than that used today.

Directed evolution has proven to be a powerful method for redesigning existing enzymes, for example with respect to changing substrate range, enantioselectivity or stability. Although engineering proteins to bind a given molecule tightly is in reach of current technologies, enzyme activity represents a much more demanding and sophisticated function, because efficient catalysis requires precise positioning of active site residues and dynamic flexibility. Indeed, our results, presented in Chapter 4, show that directed evolution on its own is not sufficient to endow an inert structure with novel activity, let alone \textit{de novo} design of active enzymes by a completely random search. Nevertheless, in 2004 an inert scaffold was endowed with triosephosphate isomerase activity through computational design and subsequently improved by directed evolution.\textsuperscript{10} This work represents a milestone towards designing an enzyme with completely new activity unprecedented in nature.

How can we create enzymes for tasks not found in nature? Screening or selection of fully randomized libraries is clearly not a viable strategy. The stepwise redesign of known sequences either rationally or by directed evolution is a possible alternative. Computational methods also have a great potential to assist enzyme engineering. Library-based approaches can be improved by increasing library size or library quality. Although it is still possible to increase experimental library sizes, ultimately, physical limits make it obvious that one should focus on increasing quality, not quantity.
To improve library quality, additional information that can be encoded in the primary sequence is needed. Such information can include using binary patterning (Chapter 2), a thermostable starting structure (Chapter 2) and reducing the redundancy in the amino acid alphabet (Chapter 3). As already noted, we are convinced that computational models will increase in importance to guide library design, and to predict functional sequences in silico. The triosephosphate isomerase example shows the enormous potential of these methods. Even if computation does not yield sequences with catalytic efficiencies as high as natural enzymes, directed evolution can then be applied to refine these leads and to tailor their properties to a wide range of applications.
6. Materials and Methods

6.1. Reagents and General Procedures

Oligonucleotides prepared by the "split and mix" technique were from Applied Molecular Evolution (San Diego, CA, USA). All other oligonucleotides were purchased from Microsynth (Balgach, Switzerland). Restriction endonucleases, Klenow fragment, T4 DNA ligase, and λ DNA were from New England Biolabs (Beverly, MA, USA). KOD Hot Start DNA Polymerase was from Novagen (Madison, WI, USA). Pfu Turbo DNA Polymerase was from Stratagene (La Jolla, CA, USA). Standard procedures were used for cloning and preparation of media. DNA sequencing was performed with the BigDye Terminator Cycle Sequencing kit from Applied Biosystems (Warrington, UK). Plasmids were isolated with the Jetquick Plasmid Miniprep Spin kit from Genomed (Löhne, Germany), and all fragments derived by PCR were confirmed by sequencing.

6.2. Vector Construction

The acceptor vectors used for library construction were isolated from the CM-deficient strain KA12 and carry a stuffer fragment to preclude artifacts due to accidental cloning of wild-type CM genes.

6.2.1. Vectors for the Libraries Described in Chapter 2 and 3

The acceptor vector for the MjCM library, pKMCMT-λ, consists of the 2,802 bp XhoI-Ndel fragment of pKECMT-W, the 964 bp Ndel-EcoRI fragment of λ DNA, and a 158 bp EcoRI-XhoI fragment encoding the 3' portion of the MjCM gene, as previously described. The acceptor vector for the EcCM and the IKEA H1 libraries, pKECMT-λ (synonymous to pECMT-λ1), is also a derivative of plasmid pKECMT-W. It was constructed by replacing the 91 bp Ndel-Nhel fragment from pECMT-W with the 999 bp Ndel-Nhel fragment from λ DNA. pECMT-W was assembled from the 212 bp Nhel-BsmAI fragment of the PCR product with template pKECMT-W and primers
AroQ1c and AroQ3t, and the 2,859 bp \textit{NheI-NotI} fragment of a \textit{BsmAI}-resistant pKECMT-W derivative. The latter was obtained by ligating the 2,313 bp \textit{BsmAI} fragment of pKECMT-W with the 776 bp \textit{BsmAI} fragment of a PCR product generated with primers Bla1c and Bla2t and template pKECMT-W.

6.2.2. Vectors for the Libraries Described in Chapter 3

For each of the two IKEA libraries H2-H3 and H3, two acceptor vectors, pECMT-Ws and pECMT-14s, were used. The plasmids pECMT-Ws and pECMT-14s were obtained by ligating the 1,246 bp \textit{EagI-XhoI} fragment of pKSS66 (the 1967 bp \textit{EaeI-XhoI} fragment from λ DNA did not give viable clones after ligation and transformation) with the 2,915 bp \textit{EagI-XhoI} fragment of pECMT-W and 14 functional IKEA H1 variants of pECMT-W (the first 14 in appendix 7.1.7.), respectively.

6.2.3. Vectors for the Libraries Described in Chapter 4

The acceptor vector for the libraries VII and XII, pECMT-stuffer (synonymous to pECMT-R20s), contains the same sequence as pECMT-W, but instead of the open reading frame encoding wild-type EcCM, pECMT-stuffer contains between its restriction sites for \textit{NdeI} and \textit{XhoI} a 1,377 bp \textit{NdeI-XhoI} stuffer fragment of pECMT-14s.
6.3. Library Construction

6.3.1. Synthesis of the FILM-DENK Libraries (Chapter 2)

The EcCM and MjCM libraries were constructed from two sets of two synthetic oligonucleotides, each purified by PAGE and about 100 nucleotides long (sequences in Appendix 7.1.1.). The oligonucleotides introduce the degenerate codons HTS (H is an equimolar mixture of A and C and T; S = C + G), which encodes Leu, Ile, Met, and Phe in a 3:1:1:1 ratio, and RAV (R = A + G; V = A + C + G), which encodes Lys, Glu, Asp, and Asn in a 2:2:1:1 ratio. The EcCM library was prepared with the 96 base long forward primer f1H1c and the 104 base long reverse primer f2H1t. The MjCM library was prepared analogously with the 99 base long forward primer RPH1CS-01 and the 97 base long reverse primer RPH1NS-02. The respective primer pairs were annealed via their complementary 3’ sequences and extended with Klenow fragment to give double-stranded DNA. The EcCM library fragment was digested with NdeI and EagI and ligated to the 2,940 bp NdeI-EagI fragment of pKECMT-λ, while the MjCM library was digested with NdeI and EcoRI and ligated to the 2,958 bp NdeI-EcoRI fragment of pKMCMT-λ. The resulting plasmids were digested with HindIII to excise the constant sequence used to anneal the oligonucleotides, and subsequently religated to generate a single HindIII recognition palindrome (AAG-CTT) encoding Lys-Leu. Aside from their different coding regions, the two library plasmid pools share the same promoter (trc), the same origin of replication (from pMB1), and the same transformation marker (a β-lactamase gene providing ampicillin resistance).

6.3.2. Synthesis of the IKEA Libraries (Chapter 3)

In IKEA libraries, only lysine plus glutamate and isoleucine plus alanine were allowed at randomized polar and apolar positions, respectively. The library design was implemented with synthetic oligonucleotides. The polar residues were specified by the degenerate RAA codon (R represents an equimolar mixture of A and G) in a 1:1 ratio. As there is no consensus codon sequence that specifies only Ile and Ala, a different strategy is required to create mixtures of these two residues. We employed the "split and
mix” technique for oligonucleotide synthesis. With this approach, the resin was split in a 3:1 ratio at positions corresponding to apolar amino acids, and the ATT or GCG codons, which respectively encode isoleucine and alanine, were synthesized separately. The resin was then remixed and the synthesis continued. The advantage of this approach is that all synthetic steps are conducted on a standard DNA synthesizer without the need for special reagents and building block libraries, such as trinucleotides or Sloning’s double-stranded tripletts. Moreover, the number, proportion, and choice of codons can be varied at will.

**Synthesis of the IKEA H1 Library (Chapter 3)**

The IKEA H1 library was constructed from the 96 base long forward primer i1H1c and the 99 base long reverse primer i2H1t. These primers were annealed via their complementary 3’ sequences and extended with Klenow fragment to make double-stranded DNA (Figure 19). By designing constant 3’ sequences that cannot form hairpin structures, problems associated with self-priming were avoided (Figure 20). The primer extension product was digested with NdeI and EagI and ligated to the 2,940 bp NdeI-EagI fragment of pKECMT-λ. The resulting plasmid pool was digested with BsmAI to excise the constant sequence used to anneal the oligonucleotides, and subsequently religated.
Two oligonucleotides, each ~100 bases

Annealing

Synthesis of second strand

Restriction digestion

Ligation

Restriction digestion

Ligation

Figure 19. Synthesis of the large IKEA H1 library module from long degenerate oligonucleotides. DNA that is removed during the synthesis is colored yellow, DNA from the (circular) acceptor plasmid is black, all other DNA is gray. Recognition sites for restriction enzymes are shown with red letters. The annealing sites of the two oligonucleotides are complementary and cannot be randomized like the library. These annealing sites were subsequently removed by a restriction enzyme that cuts outside its own recognition site. The lengths of the DNA blocks are not drawn to scale.
Figure 20. **Design of annealing region sequences.** The direction of DNA extension by the DNA polymerase is shown by arrows. Recognition sites for restriction enzymes are underlined and their cleavage sites are indicated by small arrowheads. Design A and B were successfully put to practice. Design C is very similar to A, but it results in unwanted self-priming due to hairpin formation. In design A, two fixed codons are required at the splice site because of the constrains imposed by the *HindIII* restriction site. In design B, a restriction kangaroo (*BsmAI*) was chosen that cuts outside its recognition site. The identity of this encoded amino acid can be changed easily, because *BsmAI* has no sequence preference at the site of cleavage.
Synthesis of the IKEA H2-H3 Libraries (Chapter 3)

The IKEA libraries H2-H3 and H3 were prepared with the 105 base long reverse primer i4H3t and the 87 base long forward primers i3H2c and w3H2c, respectively. The respective primer pairs (i3H2c + i4H3t or w3H2c + i4H3t) were annealed via their complementary 3' sequences and extended with Klenow fragment to give double-stranded DNA. The two primer extension products were digested with EagI and XhoI and ligated to the 2,915 bp EagI-XhoI fragments of pECMT-Ws and pECMT-14s, giving four different libraries.

6.3.3. One-Pot Synthesis of the Active Site Libraries (Chapter 4)

Two libraries, called VII and XII, were constructed from two sets of four synthetic oligonucleotides, which were purified by PAGE and contained between 58 and 76 nucleotides. The oligonucleotides for library VII contain the degenerate codon VRM (V is an equimolar mixture of G and A and C; R = G + A; M = A + C) and its reverse complement codon KYB (K = T + G; Y = C + T; B = C + T + G), which encode Arg, Lys, His, Gln, Asp, Asn, Ser and Gly with 25%, 8%, 8%, 8%, 8%, 8%, 8% and 17%, respectively. The oligonucleotides for library XII introduce the codon MDM (D = T + G + A) and its reverse complement KHK (H = A + C + T), which encode Arg, Lys, His, Gln, Asn, Ser, Ile and Leu with 25%, 8%, 8%, 8%, 8%, 8%, 17% and 17%, respectively. Library VII was prepared with the 68 base long outer forward primer b1; the 76 base long inner reverse primer d1; the 58 base long inner forward primer b2 and the 69 base long outer reverse primer d2. Library XII was prepared analogously with the 73 base long outer forward primer p1; the 76 base long inner reverse primer q1; the 68 base long inner forward primer p2 and the 74 base long outer reverse primer q2. To enable a one-pot synthesis of the libraries without purification of intermediate PCR products (Figure 16); and to avoid contamination with wild-type sequences from the PCR-template, three design features were developed. First, instead of one PCR-template encoding the whole wild-type gene, two templates were used, one encoding the first half, the other the second half of the gene. Second, the portions of the inner primers that encode the middle of the gene contain silent mutations. Third, the outer primers were employed in an access
compared to the inner primers, 0.5 μM compared to 0.1 μM, respectively. Polymerase chain reactions were performed with 0.05 U/μl KOD Hot Start DNA polymerase; 2.5 mM MgSO₄; 0.2 mM of each dNTP; 0.03 ng/μl of both templates — the 683 bp AflIII fragment of pECMT-Ws and the 632 bp BsaAI fragment of pKECMT-λ. The PCR was carried out on the PTC-200 Thermocycler (MJ Research, Waltham, MA, USA) for 50 cycles of 94°C for 30 s, 50°C to 62°C for 30 s and 72°C for 30 s. The PCR-products were digested with NdeI and XhoI, and ligated with the 2784 bp NdeI-XhoI fragment of pECMT-stuffer. Aside from their different coding regions, the two library plasmid pools share the same sequence as the libraries described above.

6.4. Selection for CM Activity

To select functional CMs from the libraries, plasmid pools were electroporated into the CM-deficient *E. coli* strain KA12/pKIMP-UAUC. Control experiments showed that electrocompetent cells of this strain typically yield 1 × 10⁸ transformants per μg of supercoiled plasmid DNA (pGEM-3Zf(+) ; Promega, Madison, WI) with large amounts of DNA (1 μg plasmid transformed). Applying the same electroporation procedure to the library plasmids prepared as described above, typically between 10⁶ and 10⁷ independent transformants per library were obtained. Larger library sizes were obtained when freshly prepared competent cells were used, compared to competent cells that were frozen and stored at -80°C. Also the growth medium used to cultivate the cells before transformation, had an effect on transformation efficiency. SOB medium without magnesium was superior to LB medium and further optimization of growth medium might improve library sizes further. After transformation, cells were washed with M9c minimal medium and plated. The total library size and unselected CM variants were obtained from colonies grown on agar plates containing phenylalanine and tyrosine after incubation for up to 7 days at 30°C (or 25°C for the experiments described in Chapter 2). Clones producing functional CMs were identified on M9c minimal medium plates lacking tyrosine (and phenylalanine for the experiments described in Chapter 2 and 3) after incubation for up to 14 days at 30°C (or 25°C for the experiments described in Chapter 2). Plasmids encoding both active and inactive CM variants were isolated and sequenced.
6.5. *In vivo* CM Activity Assay

*In vivo* CM activities of functional genes from the selection experiments were verified by retransformation into the selection strain. After retransformation of CaCl2-competent cells, they were first grown on non-selective medium. In all cases, one single colony was picked, and stroke onto selective minimal medium and (as a control) onto supplemented minimal medium. Growth on selective medium confirmed *in vivo* activities. Afterwards, the sequences of all clones were confirmed.

6.6. *In silico* Methods

The program DESIGNER scores sequences by computing an approximation to the folding free energy and uses an efficient algorithm to search through an immense number of sequences for those that minimize this free energy.125 For the *in silico* libraries, a total of 16 runs of $8 \times 10^4$ iterations were performed with the computer program DESIGNER. For all calculations, the residues at positions 46, 68, 87, 91 and 92 were allowed to explore different rotamers in order to avoid clashes with the mutated residues.

6.7. Synthesis of Designer23

Designer23 was expressed in the CM-deficient *E. coli* strain KA13 on the pET-22b-pATCH vector,148 that contained in place of its 128 bp *Ndel-XhoI* dummy fragment, the 156 bp *Eagl-XhoI* fragment from pKECMT-λ and a 132 bp *Ndel-EagI* fragment, which originated from two oligonucleotides (d1H1c 95 bases and d2H1t 79 bases) that were annealed and extended with the Klenow-fragment of DNA polymerase I.

6.8. Protein Production

The genes encoding the proteins of interest were expressed with the vector pET-22b-pATCH in the *E. coli* strain KA13.75,148 pET-22b-pPATCH is an efficient expressions plasmid that prevents translational read-throughs at the stop codon.148 In this
vector the coding sequence is under the control of the strong T7 promotor. A IPTG-inducible T7 RNA polymerase is encoded in the chromosom of strain KA13. KA13 is a RecA- and CM-deficient strain, allowing protein production without contamination with wild-type CM. In this pET vector, the gene encoding the protein of interest is inducible by IPTG and encodes a C-terminal (His)$_6$-tag, allowing purification with affinity chromatography on a resin containing chelated Ni$^{2+}$-iones.

In all cases, one single, freshly grown colony was used to inoculate each starter culture (6 ml LB medium, supplemented with 150 μg/ml sodium ampicillin) and shaken overnight at 30°C. The starter culture was used to inoculate 350 ml medium in simple 2 l Erlenmayer flasks and shaken at 30°C until the OD$_{600}$ was between 0.3 and 1. IPTG was added to a final concentration of 0.4 mM. After ca. 20 hours shaking at 25°C, the flasks were cooled down on ice and the cells were harvested at 4°C (by centrifugation at 4000×g for 10 min). The pellets were flash frozen with liquid nitrogen and stored at -80°C.

6.9. Protein Purification

Preceding protein purification, frozen cell pellets (see Chapter 6.8.) were thawed on ice, resuspended in 8 ml cold lysis buffer (containing 50 mM sodium phosphate pH = 8.0, 300 mM NaCl, 10 mM imidazole, 1 mM PMSF (Figure 21), 2 μg/ml pepstatine A, 4 μg/ml aprotinin and 1 mg/ml lysozyme) and incubated on ice for ca. 30 minutes.

![Figure 21. Protease inhibitors.](image)

(A) The serine protease inhibitor phenylmethylsulfonyl fluoride (PMSF) and (B) pepstatin A, a inhibitor of many microbial aspartic proteases.
After sonication (with 7 mm tip, 0.75 cycles, 50% amplitude, 12 times 10 s pulses followed by 10 s pauses), insoluble material was removed by centrifugation at 30,000×g for 30 min at 4°C. The supernatants were sterile filtered (through 0.22 μm Millex GP filters from Millipore, Ireland). The filtrates were loaded onto columns (with 8 mm diameter) packed with 1.4 ml Ni-NTA agarose (from Qiagen, Germany) that had been pre-equilibrated with buffer "10" (like lysis buffer, but without lysozyme and aprotinin, which has a Mw of 6511 Da). The resins were thoroughly washed, with buffer "30" and "60", (like buffer "60", but containing more imidazole, 30 mM and 60 mM, respectively). (His)$_6$-tagged proteins were eluted with buffer "250" (like buffer "10", but with 250 mM imidazole). Samples were stored at 4°C. The purification process was monitored by SDS-PAGE (using the Homogeneous 20 PhastGel from Pharmacia Biotech, now GE Healthcare, Sweden; stained with Coomassie Blue R350). As judged by SDS-PAGE, the purity of the protein samples was >95%. To concentrate the samples, to remove small molecules and to exchange the buffer, ultrafiltration was employed (with Amicon Ultra-4 devices 5000 MWCO from Millipore, Ireland).

6.10. **In vitro CM Activity Assay**

**In vitro** CM activities were measured with a well established assay under the previously published parameters (30°C, 50 mM potassium phosphate pH = 7.5, 125 μg/ml BSA). Initial rates were determined by observing the consumption of chorismate photometrically (with the UV/VIS-spectrometer Lambda 20 from Perkin Elmer) at 274 nm ($\varepsilon = 2630$ M$^{-1}$ cm$^{-1}$; for 75 μM - 300 μM chorismate) and 310 nm ($\varepsilon = 370$ M$^{-1}$ cm$^{-1}$; for 600 μM - 2500 μM chorismate) at 6 different substrate concentrations. All initial rates were corrected for the background reaction by monitoring chorismate disappearance before the addition of enzyme. Measured values were fit (with the program KaleidaGraph 3.5) to a Michaelis-Menten model to determine the kinetic constants. To allow comparison with previously published kinetic constants, protein concentrations were determined by the BCA assay with BSA as the calibration standard, rather than by UV absorption or the Bradford assay.75,128
6.11. CD Spectroscopy and Thermal Denaturation

All CD experiments were performed on an Aviv CD Spectropolarimeter, Model 202. Far-UV CD spectra were measured under the previously published parameters (20°C, PBS pH = 7.5, 5 μM protein, 0.2 cm path length, average of five wavelength scans in 0.5 nm steps, 1.5 nm bandwidth, 2 s signal averaging time).75

Also the thermal denaturation curve was measured under the previously published parameters (from 10 to 95°C in 1°C steps, PBS pH = 7.5, 1 μM protein, 1 cm path length, 222 nm, 1.5 nm bandwidth, for each step the constantly stirred sample was first equilibrated for 2 min and then the signal averaged for 1 min).75
7. Appendix

7.1. Sequences

All oligonucleotide sequences are written from the 5' end to the 3' end.
H represents an equimolar mixture of A and C and T; S = C + G; R = A + G; V = A + C + G; B = C + G + T; Y = C + T; D = A + G + T
Randomized hydrophilic positions are highlighted in blue, hydrophobic positions in red.

7.1.1. Oligonucleotides for the FILM-DENK Libraries (Chapter 2)

f1H1c for library EcCM 96 bases:
GAGCGTCTCACATATGHTSRAVRAVRAVRAVHTSHTSRAVHTSCGTRAVRAVHTSRRAVR
AVHTSRAVRAVAAAGCTTGTCACTCCCGCATCACATACCA

f2H1t for library EcCM 104 bases:
GCGAACCGCGCATGBTSADSDTTTSDGBTYSADSADBTYSAADSADSADBTYBTYAC
GBTYSADSADSADBTYSAADAGCTTGTAGTGATGCGGACTGACA

RPH1CS-01 for library MjCM 99 bases:
TATGGGAGATGGATACATATGHTSRAVRAVHTSHTSRAVHTSCGTRAVRAVHTSRRAVR
VHTSRAVRAVRAVHTSHTSAGCTTGCCTGCCCCAAGCTT

RPH1NS-02 for library MjCM 97 bases:
AAATCGGTGGAATAGGAATCCSADBTYBTYCTTSADBTYSADSADBTYBTYSADSADB
TYBTYACGTYSADSADAAGCTTGCCAGGCGAGCCAGCTT

7.1.2. Oligonucleotides for the IKEA Libraries (Chapter 3)

i1H1c for the IKEA library H1 96 bases, 8 "split+mix" codons in brackets:
GAGCGTCTCACATATG (ATT/GCG) RAARAARARAA (ATT/GCG) (ATT/GCG) RAA (ATT/GCG) CGTRAARAA (ATT/GCG) (GMW/AWA) RAA (ATT/GCG) (GMW/AWA) RAA
RAAATCGGAGACCAGCCTCGAATC
i2H1t  for the IKEA library H1  99 bases, 12 "split+mix" codons in brackets:
GCGAACCGGCCGATGTTY(AAT/CGC)(AAT/CGC)TTY(AAT/CGC)TTY(AAT/CGC)
(AAT/CGC)TTY(AAT/CGC)(AAT/CGC)(AAT/CGC)TTYTTYACGTTY(AAT/CGC)
(AAT/CGC)(AAT/CGC)TTY(AAT/CGC)GATTCGAGACCGGGGTCTCCG

i3H2c  for the IKEA library H2-H3  87 bases, 8 "split+mix" codons in brackets:
GAACATCGGCCGTTCCG(GMW/AWA)(ATT/GCG)RAACGTGAAARARAA(ATT/GCG)
(ATT/GCG)RAARAA(ATT/GCG)(ATT/GCG)RAA(ATT/GCG)(ATT/GCG)RAAG
CTCACCATCTGGATGCG

i4H3t  for the IKEA libraries H2-H3 and H3  105 bases, 15 "split+mix" codons in brackets:
TCAGTCTCGAGTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTTYTY
d1   the inner reverse primer for library VII  76 bases
CTGGACGGTGGCTCAACAGKYBGGCTTTACCAACTTCCACGGCCAGTTCACGKYBTTCT
GCCAGTAACGCTAAAG

d2   the outer reverse primer for library VII  69 bases
TCAGTCTCGAGATGTTTGTTGCAGCAACGCCTGKYBCGTCAGAAACKYBATCTTCAATGAT
GAGCTGGAAAC

p1   the outer forward primer for library XII  73 bases
GAGCGTCTCACATATGCACATCGGAAAACCCGTTACTGCGCTGDMGAAAAAMDMAGCG
CGCTGGATGAAG

p2   the inner forward primer for library XII  68 bases
GTTGAGCCACCGCTCCAGTTCTGATATTGATDMDMDCGCATMDMCTGGAAAGATTAA
TTACGCTCG

q1   the inner reverse primer for library XII  76 bases
CTGGACGTTGCTCAACAGKHKGCTTTACCKHKTTCCACGGCCAGTTCACGKHKTCT
GCCAGTAACGCTAAAG

q2   the outer reverse primer for library XII  74 bases
TCAGTCTCGAGATGTTTGTTGCAGCAACGCCTGKHKCGTCAGKKHKKHATCTTCKHKGAT
GAGCTGGAAACAGCG

7.1.4.  Oligonucleotides for Constructing Designer23

7H1c   95 bases:
GAGCGTCTCACATATGATTAAAGAAAAAAAAATTGCGGAAATTCGTAAAAAAATTGA
AAATTGATGAAGAAATCATTAAAGCGGCGGCGGAAC

7H1t   79 bases:
GCGAACCGGCCGCTCAACAGKHKGCTTTACCKHKTTCCACGGCCAGTTCACGKHKTCT
GCCAGTAACGCTAAAG

7H2t   79 bases:
GCGAACCGGCCGCTCAACAGKHKGCTTTACKGCTTCTCGCGCTTTACGCGGCTTTTTAC
GTTCCGCGCGCGCTTTAATG
7.1.5. **PCR Primers for Constructing pECMT-W**

AroQ1c 74 bases:

CGCATCGCTAGCGTCGAGGTGGGCAAAGCCAAACTGCTCTCGCATCGGCCGGTACGTGATATTGATCGTGAAC

AroQ3t 51 bases:

CGTAACGTCTCAGCCACTGGTCATTACTCGAGATGTTGTTGGAGCAAAGC

Bla1c 43 bases:

GCTACAGTCTCGCTCATAAGACAATAAACCTGATAAATGCTTC

Bla2t 41 bases:

GAACTCGTCTCTACCGGACTACCACGCTACCGGCTCCAG

7.1.6. **Functional CMs from Experiments in Chapter 2**

**EcCM variants**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLNNNENILEIREIDKLDKLLLELIIERKILKILDIKMF...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEEKILDRLRELDNMLKLELMKLRDKIIMKFLKLLK...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
<tr>
<td></td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MLNEDDLIKRLREDIEELDKLLKLLMRKLEIFIKLME...</td>
<td>MIEENELIKRFNLKLDKLLNLKLLRKKIKEIFFKIMN...</td>
<td>MNENKLLNLRDKNDKLLLELIIERKLNILKIEFKLKL...</td>
</tr>
</tbody>
</table>
MLDKENMLKLRNKIDEIDEKLKLKLIKRKEIIFEIENNIIKILE...
MFEKNNKLMEIREKIDDDLDDKLLEELLLERKKIIIKLLKFKKLNN...
MIEEDEIMNLREIIDKLDEKIILKLFFKRKNIIIFUEELLFELKLLK...
MLDKNKMPEIREIDIDTNKLNLMKRRKKLIIKIKLMKKILE...
MLEEKNFINLRRKINEIDEKLILKIEIKRELIFIDEKLKMLEN...
MLEEKKLIKLDRKIDELDEKLIEILLLDKRNLIFNNIEKILN...

<table>
<thead>
<tr>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1.7. Sequence of Designer23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Designer23 is the in silico predicted sequence (Chapter 3).

Designer23, full length amino acid sequence used for in vitro characterization:
MIKEKKAIEIRKKIEEIIKAAAERKKAIAIEAAKIIIEHRPVRDIDRERDLLELRLITLGKAHLDADHYITRLFQLIIEDSVLTTQALQLQQHLEHHHHH
Designer23, DNA sequence:

ATG ATT AAA GAA AAA AAA ATT GCG GAA ATT CGT AAA AAA ATT GAA
AAA ATT GAT GAA GAA ATC ATT AAA GCG GCG GCG GAA CGT AAA AAA
GCG GCG ATT GAA GCG GCG AAA GCG AAA ATT ATT GAA CAT CGG CCG
GTA CGT GAT ATT GAT CGT GAA CGC GAT TTG CTG GAA AGA TTA ATT
ACG CTC GGT AAA GCG CAC CAT CTG GAC GCC CAT TAC ATT ACT CGC
CTG TTC CAG CTC ATC ATT GAA GAT TCC GTA TTA ACT CAG CAG GCT
TTG CTC CAA CAA CAT CTC GAG CAC CAC CAC CAC CAC CAC TAA

7.1.8. Sequencing Primers

The following list contains the name, number of bases, estimated melting temperature (4 times the number of (G and C) plus 2 times the number of (A and T)), and the sequence in 5’ to 3’ direction of all sequencing primers used. The annealing sites of these primers are shown in the Chapter 7.1.9., page 77ff.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Length</th>
<th>Tm</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01c</td>
<td>22</td>
<td>62°C</td>
<td>ACAATTTAGGTGGCACTTTTCG</td>
</tr>
<tr>
<td>S02c</td>
<td>20</td>
<td>58°C</td>
<td>CGTTTTCCAATGATGAGCAC</td>
</tr>
<tr>
<td>S03c</td>
<td>21</td>
<td>60°C</td>
<td>TGAATGAAGCCATACCAAACG</td>
</tr>
<tr>
<td>S04c</td>
<td>21</td>
<td>62°C</td>
<td>GTGGAATTGTGAGCGGATAAC</td>
</tr>
<tr>
<td>S05t</td>
<td>21</td>
<td>60°C</td>
<td>GCCAGTGAATTGTAATACGAC</td>
</tr>
<tr>
<td>S06t</td>
<td>21</td>
<td>62°C</td>
<td>GCGTCAATACGGGATAATACC</td>
</tr>
<tr>
<td>S07c</td>
<td>20</td>
<td>60°C</td>
<td>CACTCCCGTTCTGGATAATG</td>
</tr>
<tr>
<td>S08n</td>
<td>19</td>
<td>60°C</td>
<td>CCTGACGCCAGAAGCATTG</td>
</tr>
<tr>
<td>S09n</td>
<td>19</td>
<td>58°C</td>
<td>CTCAGGCTGTAGGTATCTC</td>
</tr>
<tr>
<td>S10c</td>
<td>19</td>
<td>62°C</td>
<td>GTCAGGCAGCCATCGGAAG</td>
</tr>
<tr>
<td>S11n</td>
<td>21</td>
<td>62°C</td>
<td>GCTTCCCTTTCGGGCTTGTATA</td>
</tr>
<tr>
<td>M13(-21)</td>
<td>18</td>
<td>54°C</td>
<td>TGTTAAACGACCGGCAGT</td>
</tr>
</tbody>
</table>
7.1.9. Plasmids

Plasmid sequences were annotated with annealing sites of sequencing primer by arrows and primer name; recognition sites of important restriction enzymes by underlining and enzyme name; functional sequences by double-underlining or lines above the sequence and descriptions above the sequence. Asterisks denote stop codons. Numbering of the bases is shown on the right side and correspond to the last base on the line.

**pECMT-W**

Construction of pECMT-W is described in Chapter 6.2., page 59.

```
TTGACAGCCTTATCATCGAAGCTGCGCACCTATCACGGGAATGCTTCTGGCGTCAGG
  S10c CAGCCCATCGGAGGCCTGTTGTGATGGCTTGACGCTGAAATCACTGATA    100

ATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCC
  S07c -35 trc-promotor                                200

GACATcATAACGGTTCTGGCAAATATTCTGAAATGAGCTGTTGACA
  -10                                    300

ATCATCCGCTCGTATAATGTGAAATTGCTGGACCGGTAACAAATTTCAC
  NdeI 1 2 3 4 5 6 7 8 9 10 11 12

ACAGGAAACACGCTATATGCACTCGGAAAACCGTTATTGCACGTCGAGAG
13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
AAAATCAACGCCGTCGTTGATGAAAGGCTGCTGCTGGAACGGCG
30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45

EagI

CGAGCTAGCCGTCGAGGTGGGCAAAGCCAACTGCTCTCGCATCGGCCG
G        400

46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62
TACGTGATATTGATCGTGAACGCGATTTGCTGGAAAGATTAATTACGCTC
63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
GTAAAGCGCACCACATCGTCGCACATCAATTACGCTGCTGTCGCCG
300

80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95
XhoI

CATCAATTGAGATTGATGATGAGTTACCTGCTTAAACTCACGAGCTTGGCTCCACAACACTC
96 97 * * transcription termination

TCGAGTAATGACCAGTGCCGCTTGAATTGCCGCTAATGAGCCGAGGCT
600

TTTTTTTGAGCTCCAATTATGGAATCGATGCTTCCATATTTCAATTC S05t

ACTGGCGCTGGTTTTACCAAGGTAGTGAATGGAAAA
  M13

CCCTGGCGTTAACCACCTTAAACTGCCTTGAGCAACATCCCTCCCTTCTGCCA

GCTTGGCTAATAGCGAGAGGCCTCGCAGGCTCCCTTCACACAGTTG
  ori f1 (+)

CGAGCTTGAAATGGGACATGCGCCTGGGCGCATTAAGCGCC
782

GGCGGCGTGGGCGTTCTCACACTTGCCGAGGCGGCCC
882
```
pET-22b-pATCH

An efficient expression vector that avoids read-throughs at the stop codon.\textsuperscript{148}
7.2. Codon-Usage Table for *E. coli*

This table was adapted from the "Codon Usage Database" at www.kazusa.or.jp/codon/. Amino acids are ordered from hydrophobic to hydrophilic. Next to the amino acid name is its one letter code, followed by the three letter code, the codon and finally the percentage of each codon compared to all codons for that amino acid. For protein engineering frequent codons are preferred over rare codons.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Isoleucine  | I  | Ile | ATA | 11 % |
Valine      | V  | Val | GTG | 35 % |
Leucine     | L  | Leu | TTG | 13 % |

For example, Isoleucine is coded with the codons ATA (11%), ATT (49%), and ATC (40%).
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>One Letter</th>
<th>Three Letter</th>
<th>Codon 1</th>
<th>Codon 2</th>
<th>Codon 3</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenylalanine</td>
<td>F</td>
<td>Phe</td>
<td>TTT</td>
<td>TTC</td>
<td>58 %</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TTC</td>
<td>42 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>Cys</td>
<td>TGT</td>
<td>TGC</td>
<td>46 %</td>
<td>54 %</td>
</tr>
<tr>
<td>Methionine</td>
<td>M</td>
<td>Met</td>
<td>ATG</td>
<td></td>
<td>100 %</td>
<td></td>
</tr>
<tr>
<td>Alanine</td>
<td>A</td>
<td>Ala</td>
<td>GCC</td>
<td>GCA</td>
<td>33 %</td>
<td>23 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCA</td>
<td>23 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCT</td>
<td>18 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCC</td>
<td>26 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>G</td>
<td>Gly</td>
<td>GGG</td>
<td>GGA</td>
<td>15 %</td>
<td>13 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGA</td>
<td>13 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGT</td>
<td>35 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GGC</td>
<td>37 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tryptophane</td>
<td>W</td>
<td>Trp</td>
<td>TGG</td>
<td></td>
<td>100 %</td>
<td></td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
<td>Ser</td>
<td>AGT</td>
<td>AGC</td>
<td>16 %</td>
<td>25 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGC</td>
<td>25 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCG</td>
<td>14 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCA</td>
<td>14 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCT</td>
<td>17 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCC</td>
<td>15 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threonine</td>
<td>T</td>
<td>Thr</td>
<td>ACG</td>
<td>ACA</td>
<td>25 %</td>
<td>16 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACA</td>
<td>16 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACT</td>
<td>19 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ACC</td>
<td>40 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>Tyr</td>
<td>TAT</td>
<td>TAC</td>
<td>59 %</td>
<td>41 %</td>
</tr>
<tr>
<td>Proline</td>
<td>P</td>
<td>Pro</td>
<td>CCG</td>
<td>CCA</td>
<td>50 %</td>
<td>20 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CCA</td>
<td>20 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CCT</td>
<td>17 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CCC</td>
<td>13 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histidine</td>
<td>H</td>
<td>His</td>
<td>CAT</td>
<td>CAC</td>
<td>57 %</td>
<td>43 %</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>E</td>
<td>Glu</td>
<td>GAG</td>
<td>GAA</td>
<td>32 %</td>
<td>68 %</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>One Letter Code</td>
<td>Three Letter Code</td>
<td>Frequency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>------------------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q</td>
<td>CAG</td>
<td>66 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAA</td>
<td>34 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>D</td>
<td>GAT</td>
<td>63 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAC</td>
<td>37 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asparagine</td>
<td>N</td>
<td>AAT</td>
<td>48 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAC</td>
<td>52 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lysine</td>
<td>K</td>
<td>AAG</td>
<td>26 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAA</td>
<td>74 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
<td>AGG</td>
<td>4 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGA</td>
<td>6 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGG</td>
<td>11 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGA</td>
<td>7 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGT</td>
<td>37 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGC</td>
<td>36 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stop-codons</td>
<td>*</td>
<td>TGA</td>
<td>30 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>amber</td>
<td>TAG</td>
<td>8 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ochre</td>
<td>TAA</td>
<td>62 %</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7.3. **Image Source**

Page 40 (Monument Valley) by Moritz Zimmermann, permission is granted to copy, distribute and/or modify from [http://commons.wikimedia.org/wiki/Image:Monumentvalley.jpg](http://commons.wikimedia.org/wiki/Image:Monumentvalley.jpg)
8. References


80. Cotton RG, Gibson F. The biosynthesis of phenylalanine and tyrosine; enzymes converting chorismic acid into prephenic acid and their relationships to prephenate dehydratase and prephenate dehydrogenase. *Biochim Biophys Acta* (1965) **100**, 76-88.


105. This difference in complementation frequency cannot be attributed to the two additional N-terminal residues that were randomized in the EcCM library (Figure 5). As shown in Figure 6, significant amino acid preferences are not evident at either position, indicating that any of the available choices produces an active CM. This result is consistent with the fact that the first four amino acids in EcCM are not resolved crystallographically,(Lee *et al.* 1995) presumably due to disorder.

106. Significant covariance within the selected sequences was not observed. Detection of subtle statistical trends would presumably require a much larger sample size.
107. Notably, the importance of the residue at position 29 is not evident in an alignment of wild-type CMs.(MacBeath et al. 1998) Its critical role in EcCM was only revealed by analysis of the successful variants that were selected from the library.


Curriculum Vitae

Name: Werner Besenmatter
E-mail: besenmatter@yahoo.com
Date of birth: 12. October 1973
Place of birth: Vienna, Austria
Nationality: Austria, EU

2006 – Scientist at Novozymes A/S, Denmark

2002 – 2006 PhD in the Group of Prof. Hilvert,
Department of Chemistry and Applied Biosciences,
ETH Zurich, Switzerland

2001 – 2002 Scientific Employee at the Antibiotic Research Institute Vienna,
Biochemie GmbH, Novartis

2000 – 2001 Project Manager at the Kenya Water for Health Organization

1999 – 2000 Diplomarbeit in the Group of Prof. Hahn-Hägerdal
Institute of Applied Microbiology
Lund University, Sweden

1993 – 1999 Studies of Technical Chemistry, Biotechnology, Biochemistry
Vienna University of Technology, Austria

1988 – 1993 Federal Training and Research Institute for Industrial Chemistry,
Rosensteingasse, Vienna, Austria
Publications and Conference Presentations

Besenmatter W., Kast P. and Hilvert D. Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. 
*Proteins* (2007) 66, 500-506

Besenmatter W., Kast P. and Hilvert D.
New enzymes from combinatorial library modules. 

Besenmatter W. Plasticity of an enzyme active site. 
*oral presentation*, Protein Design and Evolution for Biocatalysis 
30.8.-1.9.2006, Greifswald, Germany

Besenmatter W. Active site engineering through designed gene-libraries. 
*oral presentation*, ENDIRPRO Meeting 2005 
9.9.-12.9.2005, Kleinwalsertal, Austria

Besenmatter W., Kast P. and Hilvert D. Thermostable enzyme significantly better as starting point for protein engineering than its homolog from *E. coli*. 
*poster presentation*, BioPerspectives 2005 
10.5.-12.5.2005, Wiesbaden, Germany

Besenmatter W. IKEA-a simple and small set of amino acids to engineer enzymes. 
*oral presentation*, Enzyme Technology and Biocatalysis 
25.11.-26.11.2004, Graz, Austria

Besenmatter W., Kast P. and Hilvert D. 
Engineering enzymes from simplified sets of amino acids. 
*poster presentation*, Biocat2004 
29.8.-1.9.2004, Hamburg, Germany