Evolutionary relationships beyond fold boundaries

Sequence and structure based exploration of two ancient superfolds

Protein chimera design by combination of related fold fragments

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Jose Arcadio Farias Rico

aus Mexiko Stadt

Tübingen
2013

Tag der mündlichen Qualifikation: 18.3.2014

Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Dr. Birte Höcker
2. Berichterstatter:	Prof. Dr. Volkmar Braun

Table of contents

List of figures and tables

Abbreviations

E. coli	*Escherichia coli*
SCOP	Structural Classification Of Proteins
ASTRAL	Compendium of sequences from structures classified in SCOP
CATH	Class Architecture Topology and Homology (classification system)
PDB	Protein Data Bank
BLAST	Basic Local Alignment Search Tool
CDD	Conserved domain database
NTM0182	N-terminal Domain of the hypothetical protein TM0182
PCR	Polymerase chain reaction
Bl21	*E. coli* expression strain
AEX	*E. coli* expression strain (ArticExpress system)
NiAC	Nickel Affinity Chromatography
IEC	Ionic Exchange Chromatography
SEC	Size Exclusion Chromatography (Gel filtration preparative)
CD	Circular dichroism
FS	Fluorescence spectroscopy
LS	Light scattering
NMR	Nuclear Magnetic Resonance
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
NCBI	National Center for Biotechnology Information
HMM	Hidden Markov Model
SAM	S-Adenosyl methionine
DUF	Domain of unknown function from PFAM
PFAM	Protein Families
kDa	Kilodaltons
UniProt	Universal protein database
IPTG	Isopropyl β-D-1-thiogalactopyranoside
PBP-like I	Periplasmic binding protein-like I fold

Summary

The comparative study of protein sequences and structures is traditionally used to better understand protein fold evolution. The insights we gain from our evolutionary analyses are applied in protein design projects. At the same time, we engineer proteins to test evolutionary assumptions; thus, we establish a feedback loop between both aspects of protein science. First, we compared Profile Hidden Markov Models, state of the art tools for homology detection, that represent all structures that adopt the $(\beta\alpha)_8$-barrel and the flavodoxin-like fold to discover an evolutionary relationship between these basic structural forms. Moreover, we located the region of the sequence space where both folds are most closely related. Having found this interface, we performed remote homologous searches and protein clustering to find sequences with intermediate features between the $(\beta\alpha)_8$-barrel and the flavodoxin-like fold.

We determined the x-ray crystal structure of one of these sequences to learn possible scenarios of fold change during the evolution of these ancestral structures. The intermediate sequence, named NTM0182, displayed features towards both folds. Moreover, and by structurally superimposing the three structures, we found classical evidences of homology among the three folds: high sequence identity over long aligned fragments. Our approach then starts by using very sensitive novel tools for homology detection (probability scores), to find intermediary links that provide classical evidences of common ancestry (high sequence identity).

Next, we extended the Profile Hidden Markov Model comparisons to include all folds classified as α/β in the Structural Classification of Proteins (SCOP). Our comparisons showed that high scoring pairwise alignments are correlated with high local structural similarities between different folds. This observation inspired us to look for interchangeable sub-domain size protein fragments, related by sequence and structure, to build chimeric proteins and mimic protein fold evolution. Our global comparisons revealed that both, the flavodoxin-like and the $(\beta\alpha)_8$-barrel folds were related to Periplasmic binding protein-like I proteins.

We then employed sequence comparisons, structural superpositions, homology modeling and computational assessments to engineer a novel chimeric protein by fusing a flavodoxin-like fold protein into a PBP-like scaffold. The chimera turned out to be a well-folded protein with native-like properties. Thus, our research allowed us to gain evolutionary insights that are applied to protein engineering. In this respect, we establish a feedback loop between fold evolution and protein engineering. We finally contribute to an emergent vision where proteins from different folds are evolutionary related.

Zusammenfassung

Die vergleichende Untersuchung von Proteinsequenzen und -strukturen wird traditionell dazu genutzt, die Evolution von Proteinfaltungen besser zu verstehen. Die Erkenntnisse, die wir aus unseren evolutionären Analysen ziehen, werden wiederum in Protein-Design-Projekten angewendet. Gleichzeitig konstruieren wir Proteine, um evolutionäre Annahmen zu testen, und schaffen so eine Feedback-Schleife zwischen den beiden Aspekten der Proteinforschung. Zunächst verglichen wir Profile von Hidden Markov Modellen, eines der modernsten Hilfsmittel zur die Homologie-Erkennung, welche alle Strukturen der $(\beta\alpha)_8$-barrel und Flavodoxin–ähnlichen Faltung repräsentieren. Dabei entdeckten wir eine evolutionäre Beziehung zwischen diesen Grundformen. Darüber hinaus identifizierten wir die Region des Sequenzraumes, in dem die beiden Faltungen besonders eng verwandt sind. Nachdem wir diese Schnittstelle gefunden hatten, führten wir eine Suche nach entfernt homologen Proteinen durch, sowie ein Protein-Clustering, um intermediäre Sequenzen zu identifizieren, die gleich weit von beiden Faltungen entfernt sind.

Wir bestimmten die Röntgenkristallstruktur von einer dieser Sequenzen, um mehr über mögliche Szenarien der Struktur-Veränderung während der Evolution ursprünglicher Faltungen zu lernen. Die intermediäre Sequenz, genannt NTM0182 , zeigte Charakteristika beider Faltungen. Darüber hinaus, und durch die Überlagerung der drei Strukturen, fanden wir klassische Hinweise für Homologie zwischen den drei Faltungen: hohe Sequenzidentität über lange Fragmente. Unser Ansatz beginnt also mit sehr empfindlichen, neuartigen Werkzeugen für die Homologieerkennung (Wahrscheinlichkeitswert), um basierend auf klassischen Homologie-Beweisen (hohe Sequenzidentität) intermediäre Strukturen zu finden.

Als nächstes haben wir Vergleiche zwischen allen Faltungen der α/β-Klasse (SCOP) durchgeführt. Unsere Untersuchungen zeigten, dass hohe Wertung bei paarweisen Alignments mit hohen lokalen strukturellen Ähnlichkeiten zwischen verschiedenen Faltungen korreliert. Diese Beobachtung hat uns inspiriert, nach auswechselbaren Proteinfragmenten zu suchen, die in Sequenz und Struktur ähnlich sind, und sich eignen um chimären Proteine zu bauen und somi die Evolution von Proteinfaltungen zu imitieren. Unsere globalen Vergleiche zeigten, daß sowohl die Flavodoxin-ähnliche als auch die $(\beta\alpha)_8$-barrel Faltung mit den periplasmatischen Bindeproteinen vom Typ I verwandt sind.

Wir haben dann Sequenzvergleiche, Strukturüberlagerungen, Homologie-Modellierung und computergestützten Berechnungen eingesetzt, um ein neuartiges chimäres Protein durch Einbauen eines Flavodoxin-ähnlichen Proteinfragments in ein PBP-Gerüst zu konstruieren. Die Chimäre erwies sich als gut gefaltetes Protein mit nativ-ähnlichen Eigenschaften. So erlaubt unsere Forschung evolutionäre Erkenntnisse, die auf Protein-Engineering angewendet werden können. In diesem Zusammenhang stellen wir fest, dass wir eine Rückkopplungsschleife zwischen Evolution und Engineering von Proteinfaltungen etabliert haben. Insgesamt tragen wir somit zu einer entstehenden Vorstellung bei, in der Proteine aus verschiedenen Faltungen in evolutionärem Zusammenhang stehen.

1. Introduction

1.1 Protein diversity and evolution

Nothing in biology makes sense except in the light of evolution (*1*). Henceforth, we study protein evolution to decipher how nature created the structural and functional diversity displayed by contemporary proteins. These molecular machines are responsible for many biological processes like DNA replication, DNA storage, protein synthesis, energy production and signal recognition among others (figure 1).

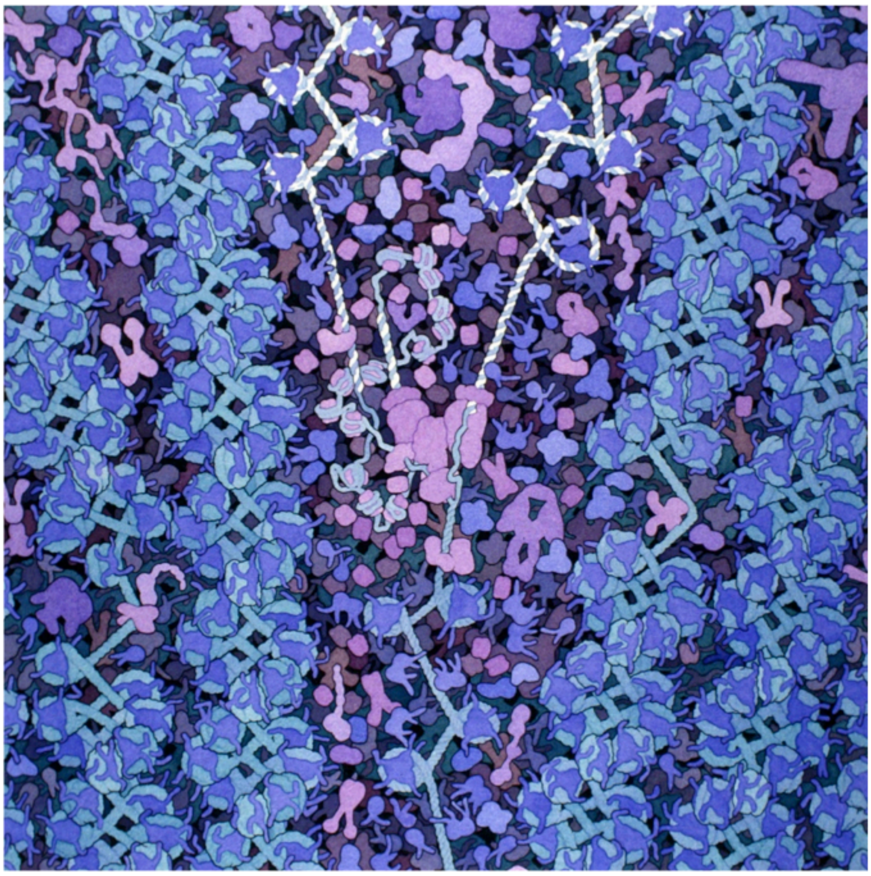


Fig. 1: DNA is replicated in the nucleus. DNA polymerase is shown at the center in purple, with a DNA strand entering from the bottom and exiting as two strands near the top. The new strands are shown in white. Chromatin fibers are shown at either site of the replication fork. (Reproduced with written permission from David S. Goodsell)

The protein structural diversity is strikingly reflected in a diversified functional range. For instance, the DNA ligase works in human cells by repairing broken pieces of DNA (*2*). Its structure is very different than the structure of bacteriorhodopsin, the biomolecule in charge of pumping protons across biological membranes during photosynthesis (*3*). The functions and structures of these two proteins are different but essential for sustaining the life of the cells (figure 2).

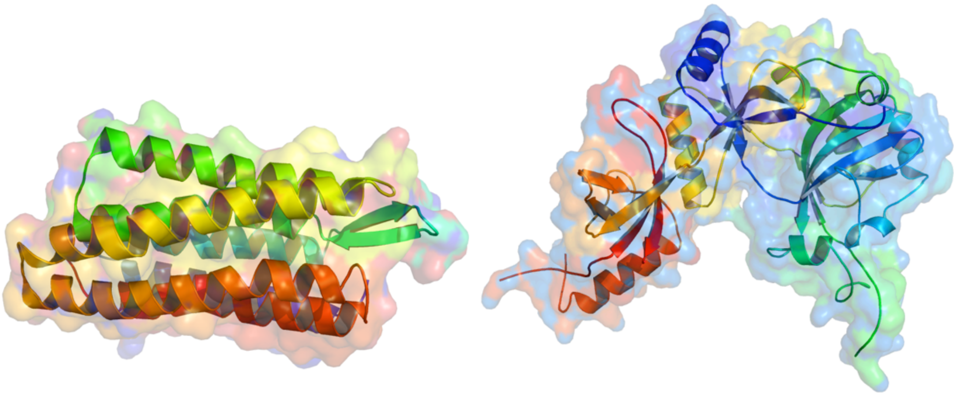


Fig. 2: Proteins are structurally diverse. The structure of bacteriorhodopsin (left) is mainly composed by alpha helices. In contrast, a mixture of alpha helices and beta strands constitutes the DNA Ligase (right).

The function and structure of proteins is defined by their amino acid sequence; therefore, it is important to understand the relationship between sequence, structure and function (*4*). It is usually understood that similar sequences will fold into similar structures, and one protein sequence will only lead to a single folded native state. The previous assumptions on the nature of proteins would then lead to a relatively straightforward relationship between sequence and structure. Nevertheless, dissimilar protein sequences can also adopt similar structures. The problem of establishing a direct correlation between sequence, structure and function represents a major challenge in the genomic and metagenomic era of biology when sequencing data is produced at an accelerated rate.

1.2 Sequence space expansion

To date, the UniProt database contains more than 46 million protein sequences (*5*), including environmental samples, from all kingdoms of life,. Scientists struggle to functionally annotate this big amount of biological data. Sequence comparisons are frequently used to infer the function of a new sequence. For instance, in the Protein Family database (Pfam) we find protein sequences grouped into families (*6*). Proteins in a family are considered to be homologous (share common ancestry) because they share significant sequence similarity, similar three-dimensional structures and functions. When a new protein sequence is added to the UniProt database, an algorithm automatically compares the protein with the families in Pfam to infer its probable function. If the new protein shares high sequence identity with a protein family of known function, it is very likely that the novel protein will have the same activity.

In Pfam we also find groups of proteins that are similar among each other and therefore are clustered in a family/group; however, they are not similar to any protein sequence with known function. These protein families are denominated Domains of Unknown Function (DUF). At the end of 2010, these families represented more than 20% of the total number of families in Pfam (*7*).

In certain cases, all the proteins in the family with unknown function display a common feature that results in the naming of this group based on a preliminary hypothesis of their probable function. For instance, in the Conserved Domain Database (CDD) (8), we found a multi-domain protein family named "B12-binding domain_like associated with radical SAM domain". The N-terminal domain of these proteins display similarities with B12-binding domains found in numerous enzymes, but they lack the signature motif Asp-X-His-X-X-Gly, which is fundamental to bind the ligand. The C-terminal domain is similar to Radical S-adenosylmethionine (SAM) proteins: these proteins generate radical species by reductive cleavage of SAM through an unusual Fe-S cluster (9). However, the function of this multi-domain group of proteins still remains unclear.

1.3 Protein structure space classification

A valuable resource frequently used to predict the structure and function of a novel protein is the Protein Data Bank (PDB). This database currently comprises almost 100,000 protein structures. The structures can assist to identify function in different ways: looking at the ligands co-crystallized with the proteins, inferring distant functional relationships, or by allowing a better definition of the domain boundaries. Nonetheless, it is necessary to build an organized picture of the structural and evolutionary relationships in the structure space, to more efficiently solve biological problems.

The databases SCOP and CATH emerged as the two main efforts to organize the structure space. SCOP (Structural Classification Of Proteins) sorts proteins into classes, folds, superfamilies and families (*8*) ; while the four major levels in CATH (*9*) are class, architecture, topology, and homology. For our work, we decided to use SCOP as the framework because it has been used as the gold standard in similar approaches.

In SCOP, four major classes constitute the top level of classification: α, β, α/β and $\alpha+\beta$. These divisions are based on the secondary structure composition of the protein structures. Subsequently, we have the fold level: protein structures that belong to the same fold have the same secondary structural elements comparably arranged in three dimensions (similar architecture) and display similar order of those elements along the path of the protein chain (similar topology). Finally, we find two more levels of classification that are based on function and sequence similarity: the superfamily and the family level. The superfamily level clusters proteins that share few identical residues, but whose structural and functional features imply common ancestry. The family level, as we previously discussed, groups proteins that share high sequence identity; therefore, common evolutionary origin for proteins classified in the same family is the most likely scenario.

1.4 Evolutionary markers in fold evolution

Random convergence to highly identical sequences, similar functions and comparable structures is very unlikely (*10*). Therefore, protein families and superfamilies are traditionally regarded as homologous groups. In contrast, two proteins may have converged to the same structure just by random physicochemical events given the small possibilities with which a polypeptide chain is folded. Thus, high sequence identity between a pair of proteins is the best indication of common ancestry. It is convenient to classify similar protein structures into folds to generate an organized view of the structure space and in this way, we discovered that the fold space is relatively small, with no more than few thousands of different basic forms (*11*). This observation leads us back to the problem of finding unrelated protein sequences (there are millions of protein sequences in the databases) that can adopt similar protein folds (a bit more than 1000 folds).

Is it possible that during the course of evolution, millions of protein sequences have diverged from a reduced number of ancestral forms? Or these sequences just converge to the reduced number of folded possibilities driven by pure physicochemical forces?

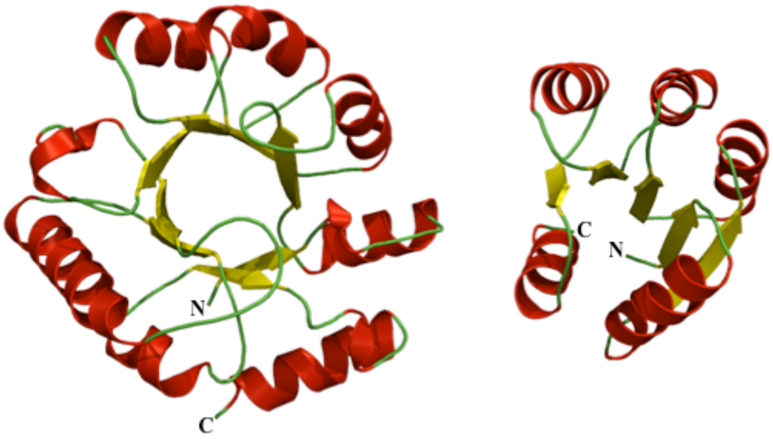


Fig. 3: Two members of the αβ class in SCOP share alternating αβ elements. The D-allulose 6-phosphate 3-epimerase from *Escherichia coli* belongs to the $(\beta\alpha)_8$-barrel fold (PDB; 3CT7, left). The response regulator CheY from *Thermotoga maritima* belongs to the flavodoxin-like fold (PDB; 1TMY, right). Beta strands are colored in yellow while alpha helices are colored in red.

Very early it was recognized that divergent evolution is capable of pushing homologous proteins to reach disparate sequences and adopt different folds (*12*). For instance, the carboxypeptidase G_2 catalytic domain from *Pseudomonas sp.* has structural similarity to the aminopeptidase from *Aeromonas proteolytica*; the similarity shows structurally aligned zinc ligands in the active site (*13*). However, the enzymes fold into different topological isomers. High local similarities in sequence and structure between proteins from different folds suggest that modern protein domains have evolved from ancient short peptide ancestors (*14*). These fragments have been denominated *antecedent domain segments* or (ADSs). The ADSs might be reflected in repetitive motifs, like the ($\beta\alpha$) elements, of the $(\beta\alpha)_8$-barrel fold (figure 3, left)

Mechanisms of protein fold change during evolution have been investigated, such as circular permutations, deletions/insertions/substitutions of whole secondary structural elements, and rearrangement of β-sheet topologies. A dramatic example is constituted by the structural comparison between bacterial luciferase and non-fluorescent flavoprotein (*15*). During the course of evolution 90 residues were deleted from the luciferase and were replaced by a single β-strand in the flavoprotein (figure 4).

1.5 Ancient protein folds in Nature

Figure 4 shows two proteins that adopt one of the most ancient protein folds, the $(\beta\alpha)_8$-barrel. This basic shape is considered among the most ancestral ones by a phylogenomic census (*16*). In this work, the usage and sharing of protein folds, in fully sequenced genomes, is employed to generate an evolutionary structured representation of the fold space. In this tree-like representation, some SCOP folds appear at the very base of the tree: P-loop containing nucleoside triphosphate hydrolase, DNA/RNA-binding 3-helical bundle, $(\beta\alpha)_8$-barrel, NAD(P)-binding Rossmann-fold domain, Ferredoxin-like, Flavodoxin-like, and Ribonuclease H-like motif. We focused our work on the $(\beta\alpha)_8$-barrel and flavodoxin-like folds. By doing this, our findings would apply to a broad range of the protein universe because these two folds are not only ancient but are also considered superfolds (*17*).

Three layers compose the flavodoxin-like fold: two α-helical layers sandwich a five-stranded β-sheet (figure 3, right). The proteins that adopt this fold perform diverse functions such as chemotaxis, binding of cofactors, signal transduction, and enzymatic activity among others (*18*) (*19*) (*20*) (*21*). Moreover, these proteins have been used as models in folding studies (*22*) (*23*) and have been ranked as ancient architectures as well (*24*). Two interesting members of these fold class are involved in completely different activities within the cell: CheY and the B12-binding domain of the Methionine synthase.

In bacteria, certain chemical signals trigger autophosphorylation of the CheA histidine kinase at a conserved histidine. Subsequently, this phosphohistidine is the substrate for CheY that catalyzes its own phosphorylation at a conserved aspartate. Then CheY-P can alter the mechanism in the flagella motor by changing the swimming speed or direction (*25*). In contrast, vitamin B12 is a common cofactor necessary to catalyze a very difficult reaction such as the removal of a methyl group. The B12-binding domain of the methionine synthase regulates the stability and reactivity of the prosthetic group by the protonation of a conserved His-Asp pair (*20*). These two examples illustrate how functionally diverse the flavodoxin-like fold is.

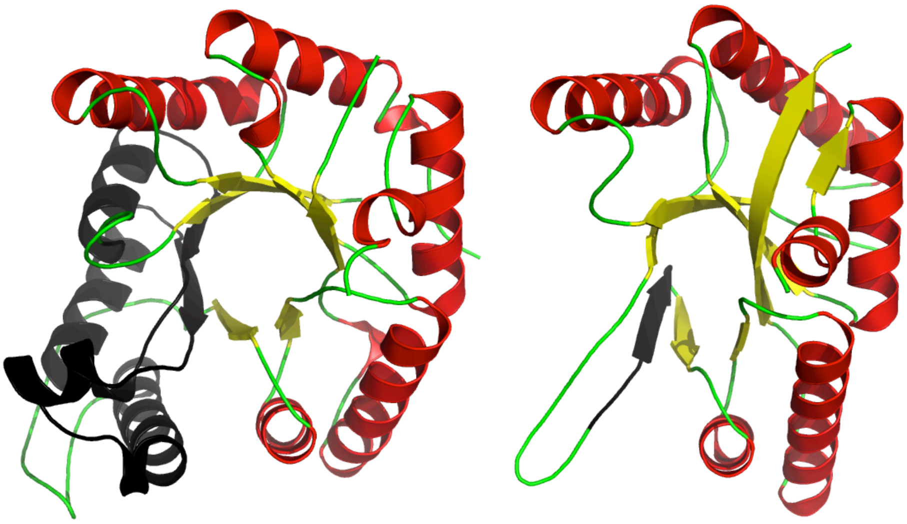


Fig. 4: Fold change in the evolution of protein structures. Bacterial luciferase (PDB-id: 1LUC, left) and nonfluorescent flavoprotein (PDB-id: 1NFP, right) The structural change is highlighted in black. The figure was adapted from (*15*).

The $(\beta\alpha)_8$-barrel fold is the most common enzyme scaffold in the Protein Data Bank (*26*); a canonical $(\beta\alpha)_8$-barrel consists of an inner ring of eight parallel β-strands that is surrounded by an outer wheel of eight α-helices (figure 4, left). Structural evidence exists that certain $(\beta\alpha)_8$-barrel proteins, like HisF and HisA, show 2-fold symmetry and they likely evolved by gene duplication and fusion (*27*). Proteins adopting this fold vary greatly in size; but 250 residues compose the typical domain. This major fold exceeds any other fold in terms of overall number and functional diversity (*28*). The functional diversity displayed by this fold class encouraged scientists to use it as scaffold for designing novel enzymes that can, for example, break a carbon-carbon bond in a non-natural substrate (*29*).

The $(\beta\alpha)_8$-barrel fold origin and evolution has been a matter of intense debate. It was not clear whether the superfamilies that adopt this fold have a common evolutionary origin. A decade ago two comprehensive works explored this possibility by a combination of sequence and structure based approaches (*30*) (*31*). At the time of this work 21 from 17 homologous superfamilies were proven to have common ancestry. One sensitive tool used for detecting homologous relationships at the superfamily level in these works was PSI-BLAST (*32*).

1.6 Methods for homology detection

PSI-BLAST takes BLAST (*32*) hits and builds a profile that is used to search a database. Therefore it is better suited to perform searches when the sequence identity of the protein pairs is lower than 30%. Nowadays, even more sensitive sequence-based tools for homology detection have been developed. For instance, HHsearch represent proteins by profile hidden Markov models (HMMs), an extension of sequence profiles that additionally trace position-specific amino acid insertion and deletion occurrences (*33*). HHsearch was recently used to explore the folds annotated in SCOP20 (sequences with less than 20% identity). Alva and co-workers (*34*) compared the HMMs of the structures to derive P-values; these values were used to cluster, using the clustering algorithm CLANS (*35*), the folds in a three-dimensional map. They showed that many of the superfamilies of one fold cluster together and that there appears to be more connectivity in fold space than expected.

The development of very sensitive tools for homology detection and their application in interesting biological problems, prompted us to reconsider our understanding of protein evolution beyond fold boundaries. For our own quest of homology beyond fold boundaries, we envisioned different starting points that are complementary.

The first point was based on an earlier observation that suggested the hypothetical homologous relationship between the $(\beta\alpha)_8$-barrel and flavodoxin-like fold (*36*). In a subsequent work, by combining sub-domain sized fragments from $(\beta\alpha)_8$-barrel and flavodoxin-like fold structures a well-folded $\beta\alpha$-barrel was built (*37*). The interface between the two folds was optimized by combining homology modeling (*38*) and computational design (*39*), yielding an eight stranded $\beta\alpha$-barrel that still retained functional properties (*40*).

1.7 Sequence based-comparisons and chimeric protein design

Based on these works we first compared protein sequences and structures to gather experimental and bioinformatic evidence for the homologous relationship between the $(\beta\alpha)_8$-barrel and flavodoxin-like folds. Once we found evidence that these two folds are related, two questions arose: Are there protein sequences with intermediary features (evolutionary bridges) between the different folds? And, can we detect other homologous folds?

We answered the first question by identifying a protein intermediate that bears similar features towards the $(\beta\alpha)_8$-barrel and the flavodoxin-like fold. We then determined the structure of this intermediate to learn more about the evolutionary path that might connect the two super folds. What is more, in our sequence-based comparisons, we not only detected homology between the $(\beta\alpha)_8$-barrel and flavodoxin-like fold. Moreover, we also found homologous fragments among many more alpha/beta folds

Among the other folds that turned out to be related with our starting folds, there was an interesting scaffold widely used in experiments of protein engineering (*41*): the Periplasmic binding protein-like I fold (PBP-like I).

We chose this scaffold to engineer a novel chimera and mimic how evolution could have generated the protein diversity we observe nowadays. The proteins that adopt the PBP-like I fold (figure 5) work as chemoreceptors and are very important members of transport systems (*42*). The fold consists of two similar intertwined lobes each composed of 3 layers (α/β/α). Each lobe displays a parallel β-sheet of 6 strands with order 213456 (*8*). The fold consists of a single superfamily of proteins.

The ligand-binding site in the PBP-like I fold is located at the hinge region that connects both lobes (*43*). Upon ligand binding there is a notable conformational change. This property has been recognized and exploited by protein engineers to build biosensors (*41*). Along these lines, the ligand specificity determination, in members of this fold, has been studied in recent works (*44*) (*45*). Biosensors capable of efficiently discriminating between similar ligands are valuable tools.

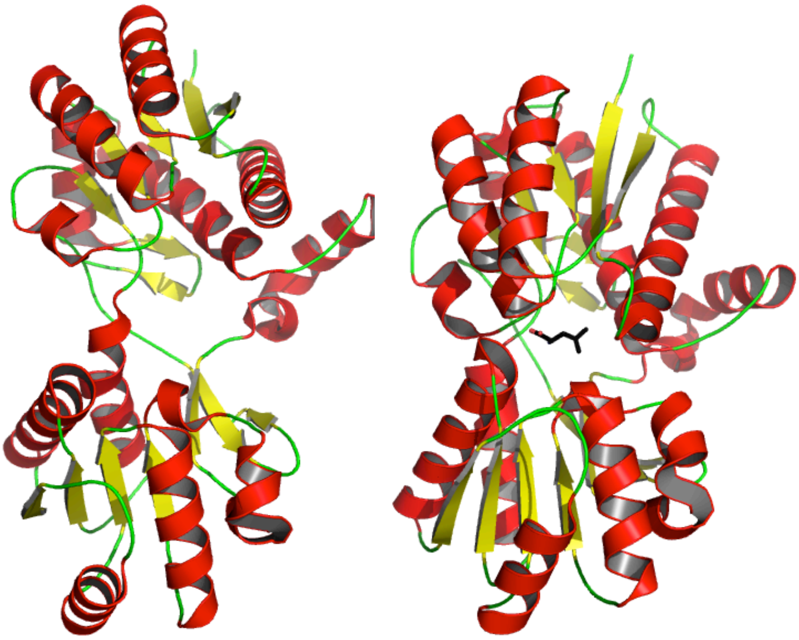


Fig. 5: The leucine-binding protein is the principal receptor for the leucine transport system in *E. coli*. The fold undergoes conformational change upon ligand binding. The open conformation (PDB-id: 1USG), depicted on the left, adopts a closed conformation when leucine (black sticks) is bound in the binding site, on the right (PDB-id: 1USK). Beta strands are colored in yellow while alpha helices are colored in red.

1.8 Origin and aim of my project

The evolution of the $(\beta/\alpha)_8$-barrel fold has been a matter of intense debate for a long time, also among my former scientific advisors. Mimicking enzyme evolution, by generating new well-folded $(\beta/\alpha)_8$-barrels from $(\beta/\alpha)_4$-half barrels (*46*), shed new light on this exciting topic. In fact, this work drew my scientific interest and was pivotal for my decision to work on this project.

The basic idea for my project was initially drafted by a publication that proposed a possible evolutionary relationship between the flavodoxin-like and $(\beta/\alpha)_8$-barrel folds (*36*). By the time we started discussing my project, another work from our group was just published describing the construction of a well-folded (β/α)-barrel by combining fragments from flavodoxin-like and $(\beta/\alpha)_8$-barrel fold structures (*37*). This experiment was a great protein engineering achievement but left opened many questions on the evolution of these two super folds. The chimeric (β/α)-barrel was designed based solely on structural information. Therefore, neither convergent nor divergent evolution could be proposed as the most likely scenario. This is why I started an exhaustive sequence based exploration of the sequence space using the powerful homology detection tools developed in house. My aim was to detect sequence-based evidence for the relationships within and between the two folds and to identify and experimentally characterize possible sequence intermediates. Our initial findings prompted us to expand the original project by comparing more folds to find related fragments that could be combined into novel chimeric constructs.

2. Materials and methods

2.1 Materials

Chemicals

Beta mercaptoethanol	Brand = Carl ROTH®
	Formula = C_2H_6OS
	MW = 78.13 g/mol
	Purity = 99%

Ethanol	Brand = Carl ROTH®
	Formula = C_2H_6O
	MW = 46.07 g/mol
	Purity 99.9%

Guanidine hydrochloride	Brand = Carl ROTH®
	Formula = $CH_5N_3 * HCL$
	MW = 95.53 g/mol
	Purity = 99.5%

Hydrochloride acid	Brand = Carl ROTH®
	Formula = HCL
	WW = 36.46 g/mol
	Purity = 37%

Imidazole	Brand = Carl ROTH®
	Formula = $C_3H_4N_2$
	MW = 68.08 g/mol
	Purity = 99%

IPTG	Brand = Sigma-Aldrich ®
	Formula = $C_9H_{18}O_5S$
	MW = 238.30 g/mol
	Purity = 99%

PEG	Brand = Carl ROTH®
	Formula = NA
	MW = 380-420 g/mol

Sodium cacodylate	Brand = Sigma-Aldrich ®
	Formula = $(CH_3)_2AsO_2Na \cdot 3H_2O$ MW = 214.03 g/mol

Sodium chloride	Brand = Merck Millipore®
	Formula = NaCl
	MW = 58.44 g/mol
	Purity = NA

Sodium hydroxide	Brand = Brand = Carl ROTH®
	Formula = NaOH
	MW = 40.01 g/mol
	Purity = 44 – 46%

Enzymes

Restriction enzymes

NdeI (Thermo scientific)
XhoI (Thermo scientific)

DNA-polymerases

Taq DNA-polymerase (Thermo scientific
Q5 DNA-polymerase (NE BioLabs$^{®}$)

DNA-ligase

T4 DNA-ligase (NE BioLabs$^{®}$)

Bacterial strains

Top10^{TM} (Invitrogen)

Escherichia coli

mrcA, Δ(mrr-hsdRMS-mcrBC), ΔlacX74, deoR, recA1, araD139Δ (ara- leu)7697, galK, rpsL, endA1, nupG

ArcticExpress^{TM} (Stratagene)

Escherichia coli

B F-ompT hsdSB (rB−m−B) dcm+ Tetr gal endA I [cpn10 cpn60 Gentr]

BL21 (DE3)

Escherichia coli

huA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5

Agarose gel electrophoresis

TAE-buffer (50x)	2 M Tris pH 8.5, 1 M HCL, acetic acid, 50 mM EDTA
Agarose gel solution	1% agarose in TAE-buffer

Buffers and solutions

Acrylamide gel electrophoresis

SDS-buffer	25mM Tris, 0.1 % SDS, 200mM Glycine
SDS-PAGE stacking gel buffer	0.5M Tris-HCl pH 6.8, 0.4% SDS
SDS-PAGE separation gel buffer	1.5M Tris-HCl pH 8.8, 0.4% SDS
SDS-PAGE staining solution	0.2% coomassie G250 and R250, 50% ethanol, 10% glacial acid
SDS-PAGE loading buffer (2x)	10% glycine, 5% β-mercaptoethanol, 2% SDS, 0.01% bromophenol blue, 1.25M Tris-HCl pH 6.8

Nickel affinity chromatography

Buffer-A	50 mM TRIS pH 7.4, 150 mM KCl, 2 mM β-mercaptoethanol
Buffer-B	50 mM TRIS pH 7.4, 150 mM KCl, 500 mM Imidazole 2 mM beta mercaptoethanol

Gel filtration-buffer

50 mM TRIS pH 7.4, 300 mM KCl, 2 mM β-mercaptoethanol

Circular dichroism

20 mM Tris pH 7.5

Refolding

R1-buffer	50mM Tris pH 7,5, 6 M guanidine hydrochloride, 300 mM KCl
R2-buffer	50mM Tris pH 7,5, 1 M guanidine hydrochloride, 300 mM KCl
R3-buffer	50mM Tris pH 7,5, 2 M guanidine hydrochloride, 300 mM KCl

Media

Luria-Bertani Medium

LB-medium	10 g peptone from bacteria, 5 g yeast extract, 5 g NaCl, water to 1 L, autoclaved

Crystallization screenings were acquired from QUIAGEN®

Instruments

Centrifuges

Bench centrifuge	Biofuge table centrifuge 5425 (Eppendorf)
Middle size centrifuge	Centrifuge 5810R (Eppendorf)
Ultracentrifuge	Avanti J-26xPI (Beckmann Coultier)

Purification columns

Nickel affinity chromatography	HisTrap HP column, 5 mL Ni Sepharose (GE Healthcare)
Preparative gel filtration	Superdex S75 GL (GE Healthcare) 320 mL V_0 Superdex S200 GL (GE Healthcare) 320 mL V_0
Analytical gel filtration	Superdex S75 GL (GE Healthcare) 25 mL V_0 Superdex S200 GL (GE Healthcare) 25 mL V_0
Ion exchange chromatography	Resource™ Q N0.520373 (6mL)

FPLC instruments

Affinity chromatography and analytical gel filtration	Aekta P900 (GE Healthcare)
	Aekta Prime (GE Healthcare)

Instruments for biophysical Characterization

Circular dichroism	J-810 CD-spectrometer (Jasco)
Fluorescence analysis	FP-6500 fluorescence-spectrometer (Jasco)

Scales

Fine scale	ALS120-4 (Kern)
Bench scale	572 (Kern)

Thermocycler

	MyCycler ThermocyclerTM (BioRad)
	T3000 Thermocycler (biometra)

Others

Nanodrop	ND-1000 Nanodrop (Peqlab ®)
pH-meter	pH211 microprocessor pH meter (Hanna Instruments ®)
UV-lamp	UV fluorescent table model ECX-20M (Peqlab ®)
Aekta superloop	Superloop 50mL (GE Healthcare)
Dialysis membranes	Spectra/Pore$^{\circledR}$ 3.5 kDa with glycerol
Shaking incubator	INOVA$^{\circledR}$ 44 incubator (New Brunswick Scientific)
Sonicator	Bandelin HD 3100 (Sonoplus)

2.2 Summary of experimental methods

We experimentally tested 25 proteins in the laboratory. 20 proteins were targeted because they showed sequence similarities towards flavodoxin-like and $(\beta/\alpha)_8$-barrel structures. Our primary goal was to characterize the structures of these proteins. For NTM0182 we determined the structure by X-ray crystallography.

Five proteins were different versions of the chimeric construct LBP-CheY. Here, our primary aim was to evaluate if the proteins were well folded. Subsequently, we aimed to generate an atomic model of the chimera. However, no crystals of enough quality to determine the structure could be obtained.

A detailed description of the methods related to NTM0182 is provided in the next section. The methods applied to the rest of the proteins are summarized in table 1.

Table 1: Proteins experimentally tested in this work. The headers of the columns are the methods applied to each protein listed in the first column. The first 19 proteins are intermediate candidates. The last five proteins are chimeric constructs (see section 8.3 from appendix for details). A symbol (✓) indicates that the method was successful and a (*X*) symbol denotes the contrary.

	Gene assembly	Purchased plasmid	PCR	BL21	ArticExpress$^{\mathrm{TM}}$	NiAC	SEC	IEC	Refolding	Crystallization
TM02			✓	✓		✓	✓			✓
CM01		✓		✓	✓		✓		✓	
CM02			✓	✓	✓		✓		✓	
SS01			✓	✓	✓				*X*	
SS02			✓	✓	✓				*X*	
SS03			✓	✓	✓				*X*	
PC01		✓		✓		✓				
PC02			✓	✓		✓				
PC03			✓	✓		✓				
PC04			✓	✓	✓	✓	✓	✓	✓	✓
PC05			✓	✓		✓				
PC06			✓	✓		✓				
HI01	✓			✓	✓				*X*	
CB01	✓			✓	✓		✓		✓	
CB02			✓	✓	✓		✓			
CT01	✓			✓		✓			*X*	
OB01		✓		✓	✓				*X*	
OB02			✓	✓					*X*	
SM01			✓	✓		✓				*X*
LBP-CheY01			✓	✓	✓	✓	✓	✓	✓	*X*
LBP-CheY02			✓	✓		✓	✓	✓	✓	*X*
LBP-CheY03			✓	✓		✓	✓	✓	✓	✓
LBP-CheY04			✓	✓		✓	✓	✓	✓	✓
LBP-CheY05			✓	✓		✓	✓	✓	✓	✓

Gene assembly: genes were assembled with overlapping primers
Purchased plasmid: genes bought from commercial supplier (*47*)
PCR: gene was obtained by PCR from genomic DNA
BL21: expression performed in *E. coli* strain BL21 (DE3)
ArticExpress: expression performed in ArticExpress system
NiAC: nickel affinity chromatography
SEC: size exclusion chromatography
IEC: Ionic exchange chromatography
Refolding: the protein was refolded from inclusion bodies
Crystallization: A crystallization screening was set up.

2.3 Cloning methods

21 variants of intermediate sequences and 5 versions of the LBP-CheY chimera were experimentally tested in this work. All constructs were cloned in the vector pET-21a-d(+) from Novagen®. The restriction sites used were *XhoI* (site 158) and *NdeI* (site 238). The list of sequences, primers and special remarks about the cloning are provided in appendix section 8.3. In the following tables we describe the common reactions for PCR amplification, digestion, ligation, colony PCR and sequencing for the cloned variants:

Table 2: Polymerase chain reaction mix

Reagent	Volume	Final concentration
Q5® reaction buffer (NE BioLabs®)	10	1x
dNTPs (Eurogentec®)	1	10 mM
Forward primer	2	10 mM
Reverse primer	2	10 m0
DNA template (plasmid or genomic DNA)	0.5	~50 ng
Q5® high-fidelity DNA polymerases (NE BioLabs® $_{\text{inc}}$)	1	2u/µl
H_2O	32	
Total volume	50	

Table 3: Polymerase chain reaction cycles

Degrees (Celsius)	Time	Cycles
95	60 sec.	1
95 65 (variable temperature) 72	10 sec. 30 sec. 20 sec.	30
72	5 min.	1

Table 4: Digestion of plasmids and PCR products.

Reagent	Volume µl	Final concentration
NdeI (Thermo scientific) 100% activity	1	1.0 U
XhoI (Thermo scientific) 50-100% activity	1	1.0 U
DNA template	40	~ 1 µg
Buffer O (orange)	5	1x
H_2O	3	
Total volume	50	

The DNA is incubated overnight at 37 °C.

Table 5: DNA ligation reaction.

Reagent	Volume µl	Final concentration
T4 ligase (NE BioLabs®)	1.5	1.5 U
DNA insert (variable)	2 (variable)	~ 50 ng. (variable)
DNA vector (variable)	10 (variable)	~ 500 ng. (variable)
T4 ligase buffer (NE BioLabs®)	2	1x
H_2O	4.5	
Total volume	20	

The DNA is incubated overnight at 4 °C.

Table 6: Polymerase Chain Reaction from colony.

Reagent	Volume µl	Final concentration
10 x Taq Pol buffer	2	1x
dNTPs (Eurogentec®)	2	200µM (each Nucleotide)
Forward primer	1	1 m
Reverse primer	1	1 m
Taq DNA polymerase	2	2u/µl
H_2O		
Total volume	20	

Table 7: DNA sequencing reaction

Reagent	Volume µl	Final concentration
DNA plasmid	1	50 ng.
BDT-Mix	0.5	1x
Sequencing primer	1	1 m
Sequencing buffer	2	1x
H_2O	5.5	
Total volume	10	

Gene assembly of intermediate candidates

Three intermediate candidates were synthetized by assembling alternating primers. The method employed in this work is a variation of the methodology outlined in the following publication: " Simplified gene synthesis: A one-step approach to PCR-based gene construction" (*48*). Basically, it is necessary to try different oligonucleotides concentrations but always adding the external set of primers at much higher (10x) concentration than the internal set. By this, we ensure that the full product is amplified.

Gene purchasing

Three intermediate candidates were purchased from "Mr. Gene" a company specialized in gene synthesis. Mr. Gene was founded in late 2006 and later was bought by life technologies® (*47*). The specialized service was renamed as: GeneArt®. The constructs can be optimized for *E. coli* expression and restriction sites are also engineered. Upon digestion it is possible to sub-clone the construct in an expression vector.

PCR assembly of LBP-CheY chimeras

The Leucine Binding Protein, from *Escherichia coli*, was assembled from genomic DNA using the following primers in a standard PCR reaction:

Primers: "FA_fwd" and "FC_K_rev"
Template: Genomic DNA

The sequence annotated in the NCBI database had a mutation when compared with the crystal structure sequence. We cloned the protein following the sequence associated with the PDB structure (with the K instead of the T). See appendix section 8.3 for details. The region from the Response regulator CheY (*Thermotoga maritima*) that was inserted in the LB-CheY chimera was amplified from a plasmid kindly donated by Dr. Simone Eisenbeis using the following primers:

Primers: "FB_fwd" and "FB_rev"
Template: Plasmid pET21-CheYWT

LBP-CheY-01 assembly

Five reactions were needed to build the chimeric construct:

Reaction 1: fragment A (N-terminus region LBP)

Primers: "FA_fwd" and "FA_rev"
Template: pET21-LBP

Reaction 2: fragment B (CheY region)

Primers: "FB_fwd" and "FB_rev"
Template: pET21-CheYWT

Reaction 3: fragment C (C-terminus region LBP)

Primers: "FC_fwd" and "FC_K_rev"
Template: pET21-LBP

Reaction 4: join fragments A and B.

Primers: "FA_fwd" and "FB_rev"
Templates: fragments A and B

Reaction 5: join fragments AB and C

Primers: "FA_fwd" and "FC_K_rev"
Templates: fragments AB and C

LBP-CheY-02 assembly

The construction of chimera LBP-Che-02 was identical to the method followed for LBP-Che-01 but the fragments (reactions2 and 3) B and C were built using different primers to introduce the extra K:

Reaction 2: fragment B (CheY region)

Primers: "FB_fwd" and "FB_K_rev"
Template: pET21-CheYWT

Reaction 3: fragment C (C-terminus region LBP)

Primers: "FC_K_fwd" and "FC_K_rev"
Template: pET21-LBP

LBP-CheY-03 assembly

The chimera LBP-Che-03 only had the 6X His Tag removed by PCR:

Primers: "FA_fwd" and "chimera-lbp-chey-c-ter-A344>K"
Template: pET-LBP-CheY-02

LBP-CheY-04 assembly

New primers were used to change the residues at the interfaces of the fragments from both folds.

Reaction 1: fragment A (N-terminus region LBP)

Primers: "FA_fwd" and "FA_I_rev" (change this in list of appendix)
Template: pET-LBP-CheY-02

Reaction 2: fragment B (CheY region)

Primers: "F_B_I_fwd" and "F_B_MTV_rev"
Template: pET-LBP-CheY-03

Reaction 3: fragment C (C-terminus region LBP)

Primers: "F_C_MTV_fwd" and "FC_K_rev"
Template: pET21-LBP

The rest of the reactions are the same than previous versions

LBP-CheY-05 assembly

Three reactions are needed.

Reaction 1: fragment A (N-terminus region LBP)

Primers: "FA_fwd" and "RV_mid-LBPCHEY_version10"
Template: pET21-LBP-CheY-03

Reaction 2: fragment LBP-C

Primers: "FW_mid-LBPCHEY_version10" and "chimera-lbp-chey-c-ter-A344>K"
Template: pET21-LBP-CheY-03

Reaction 3: join both fragments

Primers: "FA_fwd" and "chimera-lbp-chey-c-ter-A344>K"
Template: Fragment A and LBP-C

Gel extraction.

All digested PCR fragments and vectors were loaded onto an agarose gel 1%. Desired bands were cut with a sterile razorblade. The purification was performed according to the protocol in *QIAquick gel extraction kit* of QIAGEN®. Elution volumes ranged from 500 ml to 50 ml according to the concentration observed on the gel.

Transformation

The tubes with competent cells were thaw on ice. We add ~50ng (ligation: 10-15µl) of DNA to the cells and let it sit on ice for ~15min. Heat shock the cells for 45sec. (ArcticExpress^{TM} cells: 20sec.) at 42°C. Put them back on ice for ~10 min. Add 900 µL of LB media (no antibiotics) to the cells and incubate them for 1h at 37°C. Plate out 100µL on LB-agar plates (ligation: spin down the cells at 4000rpm for 10 min and resuspend the pellet in a little bit of supernatant; plate out all the solution).

2.4 Heterologous Expression

The protocol outlined here has to be implemented after transformation:

- Pick some colonies from a freshly transformed plate
- Start an overday culture using Luria-Broth (LB) media (5ml)
- Start and overnight culture (+AB) 50mL LB
- The next day inoculate 2L LB (+AB) with the 20mL of O/N-culture
- Grow at 37°C (this can be also at 20, 25 or 30 °C) until the OD_{600} is ~0,7
- Induce the cells by adding isopropyl-β-thiogalactoside (IPTG) to a final concentration of 1 mM
- Allow growth for 4h at 37°C or longer at lower temperatures (up to 24 hrs. at 20 °C)
- Harvest the cells by centrifugation (4000rpm, 15 min, 4°C)
- Take off the supernatant and resuspend the pellet in 20-50 mL buffer
- Centrifuge again (4000 rpm, 15 min, 4°C)
- Resuspend the pellet in 20-50 mL buffer and add protease inhibitors (Protease-Inhibitor Mix HP, Serva).
- Sonicate the cell suspension (40% amplitude, 10 min [0.1sec pulse on, 0.9sec pulse off] or 5 min [0.2sec pulse on, 0.8sec pulse off])
- The resulting homogenate should be centrifuged (18000 rpm, 40 min, 4°C).
- Filter the supernatant homogenate with 0.22 μM filter before proceeding to any purification protocol

ArcticExpress heterologous expression system

The overall protocol is similar to the BL21 (DE3) expression but with the following differences:

- Transform as previously indicated (but pulse of 20 sec.) the competent cells with the protein expression plasmid, using a 37°C cultivation temperature.
- Pick several transformants and grow overnight cultures in medium containing gentamycin and antibiotic for selection of the expression plasmid at 37°C.
- Grow the cells without antibiotic selection for 3 hours at 30°C.
- Induce expression of the protein with IPTG at 10–13°C, and continue growth after induction at 10–13°C for 24 hours (or more)
- Analyze protein expression in induced cultures and non-induced controls by SDS-PAGE.

Refolding

- Take the pellet after sonication (coming from an expression protocol) and resuspend it in 10 mL 6M guanidine hydrochloride (GdHCl)
- Let it stand for 60 min at 4°C (cold room)
- Add 10mL 1M GdHCl and mix it by swiveling it gently (this is important)
- Let it stand for 60 min at 4°C
- Centrifuge (18000rpm, 60 min, 4°C), afterwards the protein is in the supernatant.
- Take the supernatant and add 2M GdHCl to a final volume of 50mL (a falcon tube can be used)
- Dialyze against 3 x 5L of buffer and continue with further purification steps. Diverse additives could be used during the refolding procedure.

2.5 Purification protocols

Table 8: Buffers employed with the proteins experimentally tested in this work.

Buffer	Method	Contents	Protein variants
A	NiAC	50 mM TRIS pH 7.5 150 mM KCl 2 mM β-mercaptoethanol 50 mM NaCl 10 mM Tris pH 7.6	All variants (Except LBP-CheY variants) LBP-CheY variants
B	NiAC	1 M imidazole 50 mM TRIS pH 7.5 150 mM KCl 2 mM β-mercaptoethanol 50 mM NaCl 10 mM Tris pH 7.6 500 mM Imidazole	All variants (Except LBP-CheY variants) LBP-CheY variants
SEC	SEC	50 mM TRIS 300 mM KCl pH 7.5 2 mM β-mercaptoethanol	TM01, TM02, CM01, CM02 SM01 LBP-CheY variants
Sample	CD spectra	20 mM Tris pH 7.5 10 mM Tris pH 7.4	TM01 CM01 LBP-CheY variants
Lysis	BL21 (DE3) Expression AEX expression	50 mM TRIS 150 mM KCl pH 7.5	All variants
Sample	Fluorescence spectroscopy (FS)	20 mM Tris pH 7.5 100 KCl 10 mM Tris pH 7.6 300 mM NaCl	TM01 LBP-CheY variants
Potassium phosphate buffer A	IEC	KP 100 mM pH 6.8	PC04
Potassium phosphate buffer B	IEC	KP 100 mM KCl 500 mM pH 6.8	PC04
SEC	SEC	KP 100 mM KCl 300 mM pH 6.8	PC04

Buffer	**Method**	**Contents**	**Variants**
A	NiAC, SEC	10 mM β-mercaptoethanol 50 mM HEPES 7.5 10 mM Cl_2Mg 300 mM KCl	PC04
B	NiAC	10 mM β-mercaptoethanol 50 mM HEPES 7.5 500 mM Imidazole 10 mM Cl_2Mg 300 mM KCl	PC04
Sample	Light Scattering	50 mM TRIS 300 mM KCl pH 7.5	TM01

NiAC: Nickel Affinity Chromatography

We follow the protocol published elsewhere (*49*) First we lyse the cells that are expressing the tagged protein by sonication on ice or French press. We used different lysis buffers for different variants (table 8). Approximately 3–5 mL of loading buffer should be used per gram (wet weight) of cells. Keep the lysate as cold as possible to minimize protein degradation.

The protein sample was loaded onto a 5 mL nickel column (*HisTrap HP* Sepharose GE Healthcare) previously equilibrated with 50 mL of buffer A. The target protein was eluted with an imidazole gradient. We collected the fractions that showed the highest concentration of recombinant protein. The samples must be analyzed by SDS-PAGE.

IEC: Ionic exchange chromatography

We applied a modify version of the protocol outlined elsewhere (*50*). The protein sample should not contain any salt. The sample can be dialyzed against the equilibration buffer for the column. The protein sample was then loaded onto a 6 mL Resource^{TM} column via the injection loop. Later, the column is washed with at least 5 column volumes) of starting buffer or until baseline is reached. The protein is then eluted with 10–15 column volumes of a salt gradient.

SEC: Size exclusion chromatography

We used the method described elsewhere (*51*). The sample has to be pre-purified with another method. It is not recommended to inject crude extract into a gel filtration column. Importantly, the sample should be dialyzed against gel filtration buffer and it must be filtered through a 0.22-µm protein-compatible filter. It is necessary to avoid air bubbles that will end up in the top of the column. The sample is loaded via an injection loop and eluted with at least one column volume of buffer.

2.6 NTM0182 methods

Since we performed a full experimental characterization of NTM0182, that includes structure determination, we outline detailed methods related with this variant. NTM0182 carries a His_6-tag at its C-terminus; it was produced in *E. coli* BL21 (DE3). The cells were grown at 37°C in Luria-Broth supplemented with 100 μg/mL ampicillin for maintenance of the plasmid. At an OD_{600} of 0.6, adding isopropyl-β-thiogalactoside to a final concentration of 1 mM induced expression, and growth was allowed for another 15 h. NTM0182 was mainly found in the soluble fraction of the cell extract and purified from this fraction. Cells were harvested by centrifugation, washed with 50 mM Tris (pH 7.5), 150 mM KCl, and centrifuged again.

The cells were resuspend in 20 mL of same buffer, and protease inhibitor was added in standard 1x concentration (Protease-Inhibitor Mix HP, Serva). The cells were lysed by sonication (Branson Sonifier W-250, 6 × 1.0 min, Output 5, 50% pulse, on ice), and the resulting homogenate was centrifuged (18000 rpm, 40 min, 4°C). The supernatant was filtered and loaded onto a NiNTA column (Amersham Pharmacia) equilibrated with 50 mM Tris (pH 7.5), 150 mM KCl. The protein was eluted with an increasing concentration of imidazole.

Fractions with the highest content of protein were dialyzed extensively against 50 mM Tris (pH 7.5), 300 mM KCl and then loaded onto a Superdex 75 HiLoad 26/60 column (320 mL, Amersham Pharmacia), which was equilibrated with the same buffer. Elution was performed at a flow rate of 2.0 mL/min. The protein eluted mainly in two well-differentiated peaks, which were collected and treated independently.

Analytical Methods

Purification of the proteins was evaluated by electrophoresis on 15% polyacrylamide gels, using the system of Lämmli (*52*) and staining with Coomassie blue. Protein concentrations were determined by using molar extinction coefficients calculated from the amino acid sequence. Analytical gel filtration was performed by using a calibrated Superdex 75 10/300 GL column (Amersham Pharmacia) and was coupled to a light scattering device for subsequent analysis.

Multiangle static laser light-scattering experiments (MALLS) were done online with analytical size-exclusion chromatography using miniDAWN TREOS and Optilab rEX instruments (Wyatt Technologies) and the associated software (AstraV) for molecular weight determination; the method was published elsewhere (*53*). The protein (0.3 mg/ml) was eluted at a flow rate of 0.8 ml/min in 50 mM Tris, 300 mM KCl (pH 7.5) and the apparent size of the two peaks was analyzed both by size exclusion and dynamic light scattering. CD spectra were recorded with a JASCO model J-810 spectropolarimeter. Fluorescence measurements were carried out with a JASCO FP-6500 spectrofluorometer. The measurements were performed in 50 mM Tris and 300 mM KCl (pH 7.5) at room temperature.

Crystallization trials (8 screenings, 96-well plates) were set for both oligomerization probes from native NTM0182. No crystals were obtained for the monomeric NTM0182. Small needle-like crystals were obtained with the dimeric NTM1082 and further refined by the hanging drop vapor diffusion method at 18°C. Drops contained 1.5 µl of the protein solution (11.37 mg/mL) mixed with 1.5 µl of 0.1 M HEPES at pH 7.5 with 1.2 M ammonium sulfate and 0.3 M NaCl, and were equilibrated against 500 µl of reservoir buffer. After short transfer into crystallization buffer with 25% glycerol, the crystals were flash frozen in liquid nitrogen. Soaking of crystals was performed with Potassium tetra-cyano-platinate (II) hydrate [K2Pt(CN)4 • xH2O; Mw: 377.36]. Derivative crystals were washed using crystallization buffer with 25% glycerol; afterwards the crystals were flash frozen in liquid nitrogen. Data was collected at the synchrotron beamline PXII (Swiss Light Source, Villigen PSI) at 100K, and 0.5 oscillation degrees (images) were recorded on a PILATUS 6M 500-mm detector. Native and derivative platinum crystals were measured at a wavelength of 1.0000 and 1.0698 Å, respectively. Data were indexed, integrated, and scaled with XDS and XSCALE and converted with XDSCONV (*54*)

Heavy atom detection in derivative crystals was done using SHELX (*55*). Two sites were unambiguously detected via a SAD protocol (*56*). The unit cell dimensions between the native and two derivative crystals were too big to perform single isomorphous replacement. Heavy atom model refining (on the two detected platinum sites), phasing and density modification (using the programs DM and SOLOMON) was performed using SHARP (*57*). We combined the data from the two platinum soaked crystals in a single SAD experiment (via the SHARP interface) in order to achieve enough phasing power to produce an initial experimental map. After density modification and solvent flipping the map showed clear density for α-helices and β-sheets. An initial model was built using a combination of Phenix (*58*) (to find helices and strands) and Buccaneer (*59*) (for fast chain tracing) in command line version. Non Crystallographic Symmetry (NCS) was used to build six chains into the asymmetric unit taking as starting molecule chain A. Model building was performed with the program COOT (*60*).

Initial refinement was done under NCS, Ramachandran, and experimental phase restraints with Phenix refine. Intermediate refinement rounds were performed with: ZYX coordinates, group B factors (a single b-factor per residue), TLS parameters and anomalous groups. Final refinement resulted in R_{cryst} and R_{free} values of 24.6% and 28.8%, respectively.

2.7 Bioinformatic analysis

Sequence comparisons.

The sequence based comparisons between the $(\beta\alpha)_8$-barrel fold structures (SCOP id c.1) and flavodoxin-like fold structures (SCOP id c.23) were performed with HHsearch (*33*) The software provides a library of Hidden Markov Models representing all the structures in SCOP70 and SCOP95 (Structural Classification of Proteins and ASTRAL releases 1.75 and 1.75B (the later from January 2013). We extracted all models representing all α/β fold profiles and used them as queries to search the entire SCOP95 database of models. Afterwards we filtered the $(\beta\alpha)_8$-barrel and flavodoxin-like fold outputs (890 profiles).

The comparisons for all the alpha folds were done with a previous version of SCOP (SCOP70). Only the comparisons for (βα)$_8$-barrel fold structures (SCOP id c.1) and flavodoxin-like fold structures (SCOP id c.23) were performed using SCOP95.

We used default parameters; however, we did not score secondary structure alignment to avoid biases introduced by the highly similar secondary structure content of the folds. High probability hits were recorded at three arbitrary cutoffs: 100-80 b) 79.9-60 c) 59.9-40

Identification of the intermediate sequences

HHsenser was used to perform remote homologous searches (parameters: Database= *nr* + environmental, Extension of the seed=50, PSI-BLAST E-value threshold=1e-3, Minimal coverage PSI-BLAST hits=20, Use clustered database=No, Terminate search=5000 sequences found, Prescreen for structural domains=No, using as queries the sequences from several flavodoxin-like (e.g. PDB's: 1I9C, 2YXB, 1TMY) and (βα)$_8$-barrel fold structures (e.g. 3IGS, 1THF and 1ZFJ). From each search around 3500 homologous hits were collected. The hits were merged and fed to the clustering program CLANS with which 5000 rounds of clustering were performed. Finally, a visual inspection of the cluster map allowed the identification of sequences that link clusters from different folds, or that were clearly located between them, at an astringent P-value (1.0E-10).

Structural superpositions guided by profile-profile sequence alignments: HHpred searches were started with default parameters (web version) using the sequences of the aligned structures as queries in order to generate profile-profile pairwise sequence alignments among these proteins. The sequence alignments were then used as guides to structurally superimpose the three proteins using the software PDBefold (*61*)

Homology modeling and Rosetta relaxation runs of LBP-CheY chimeras

We submitted the sequence of the chimeric proteins to the HHpred (*62*) server to obtain an alignment and prepare the input file for the Modeller (*38*) job. We then edit the alignment as shown in appendix (8.2 section). We ran modeler in the web server to produce the homology model.

This model was used as input for Rosetta relax (*63*). We also use the PDB structures from LBP (1USK and 1USG) and CheY (1TMY) to perform relaxation and get initial energies per residue. Per every variant tried in this work we generated 100 structures, and averaged the top best in terms of Rosetta energy units. To perform the above mentioned protocol we wrote a python script that process the output file "score.sc" taking the column that corresponds to the total score per structure. The total score per structure is the total contribution of energy per residue. The script outputs the top five models from the run in terms of *rosetta* energy units. We afterwards averaged the energy of the top five structures. The idea is to compare the energy per residue of the parental structures and the respective energies in the new chimeric context. Residues with unusually high energies might indicate problematic areas that can be visualized. This could represent a problematic area that may prevent the protein to reach a native-like folded conformation.

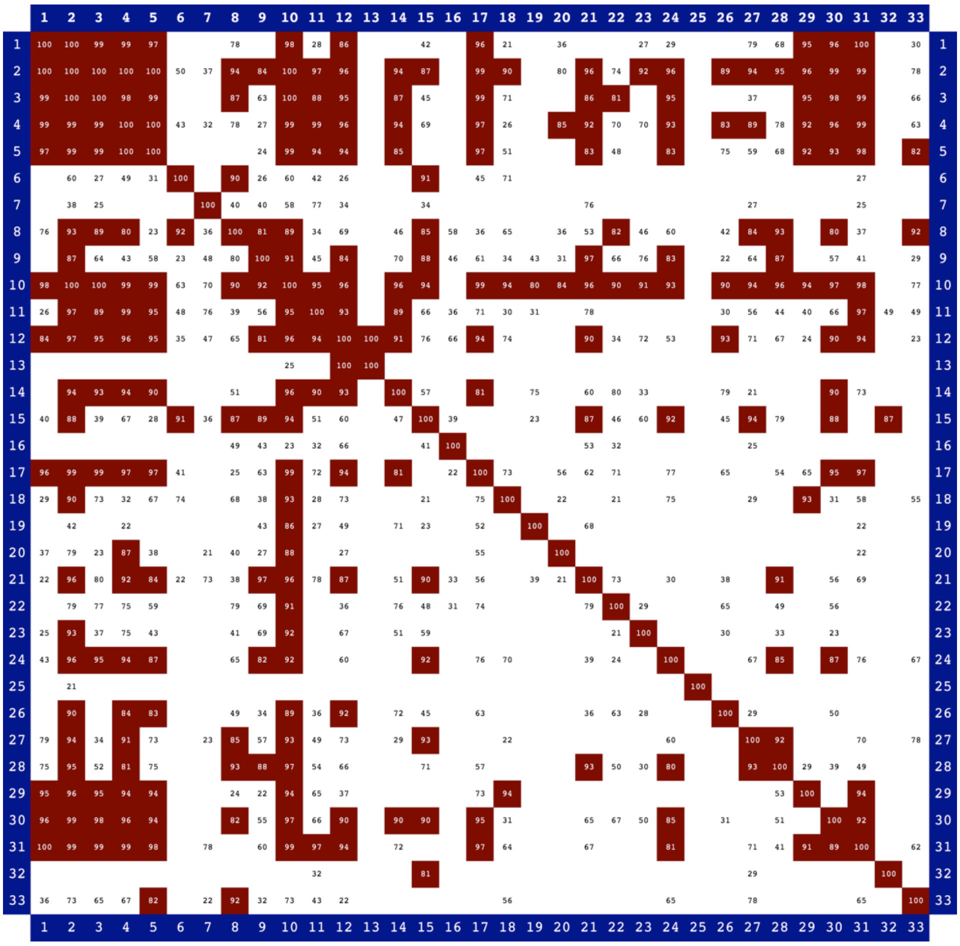
3. Results

3.1 $(\beta\alpha)_8$-barrel fold evolutionary relationships

The common evolutionary origin of the $(\beta\alpha)_8$-barrel fold has been a matter of intense debate over the past decades. Our first goal was to generate an updated overview of the relationships that could be detected with state of the art tools for homology detection. Therefore we performed pairwise profile-profile comparisons using the sequences of the structures classified in the Structural Classification Of Proteins (SCOP). We worked at the superfamily level of classification given that members of the same superfamily are believed to share a common evolutionary origin. There are defined 33 homologous superfamilies that adopt the $(\beta\alpha)_8$-barrel fold. Previously, the hypothetical monophyletic origin of the fold has been extensively studied in two works (*30, 31*).

Now, through profile Hidden Markov Model (HMM) comparisons, we not only confirm previous findings, but we also found evidences of common ancestry among additional $(\beta\alpha)_8$-barrel fold superfamilies. The sequence based comparisons between the $(\beta\alpha)_8$-barrel fold structures (SCOP id c.1) and flavodoxin-like fold structures (SCOP id c.23) were performed with HHsearch. We used default parameters; however, we did not score secondary structure alignments to avoid biases introduced by the similar secondary structure content of both folds. Hits were recorded from 100 to 20 percent HHsearch probability. The score is as Bayesian posterior probability, which represents the level of certainty to a potential outcome. This kind of likelihood is a different interpretation of the concept of chance and belongs to the category of evidential probabilities. The frequentist probability in contrast, defines an event's probability as the limit of its relative frequency in a large number of trials. An event with Bayesian probability of 0.6 (or 60%) should be interpreted as stating "with confidence 60%, this event contains the true outcome", whereas a frequentist interpretation would view it as stating "over 100 trials, we should observe event X approximately 60 times (*64*)".

In agreement with previous studies, 30 of 33 SCOP homologous superfamilies showed pairwise connections at a better probability of 80 percent; see figures 6, 9 and 10 for details. The probability graph in figure 6 is not perfectly symmetrical due to the intrinsic properties of the HMMs. For any given pairwise profile comparison between two superfamilies, the recorded probabilities lie in the range of 20 to 100 percent and the hits can be bidirectional or unidirectional. It is intuitively assumed that 80% bidirectional probability hits will more likely indicate common ancestry than 20 % unidirectional hits.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1	100	100	99	99	97			78		98	28	86			42		96	21		36			27	29			79	68	95	96	100		30
2	100	100	100	100	100	50	37	94	84	100	97	96		94	87		99	90		80	96	74	92	96		89	94	95	96	99	99		78
3	99	100	100	98	99			87	63	100	88	95		87	45		99	71			86	81		95			37		95	98	99		66
4	99	99	99	100	100	43	32	78	27	99	99	96		94	69		97	26		85	92	70	70	93		83	89	78	92	96	99		63
5	97	99	99	100	100				24	99	94	94		85			97	51			83	48		83		75	59	68	92	93	98		82
6		60	27	49	31	100		90	26	60	42	26			91		45	71													27		
7		38	25				100	40	40	58	77	34			34						76						27				25		
8	76	93	89	80	23	92	36	100	81	89	34	69		46	85	58	36	65		36	53	82	46	60		42	84	93		80	37		92
9		87	64	43	58	23	48	80	100	91	45	84		70	88	46	61	34	43	31	97	66	76	83		22	64	87		57	41		29
10	98	100	100	99	99	63	70	90	92	100	95	96		96	94		99	94	80	84	96	90	91	93		90	94	96	94	97	98		77
11	26	97	89	99	95	48	76	39	56	95	100	93		89	66	36	71	30	31		78					30	56	44	40	66	97	49	49
12	84	97	95	96	95	35	47	65	81	96	94	100	100	91	76	66	94	74			90	34	72	53		93	71	67	24	90	94		23
13										25		100	100																				
14		94	93	94	90			51		96	90	93		100	57		81		75		60	80	33			79	21			90	73		
15	40	88	39	67	28	91	36	87	89	94	51	60		47	100	39			23		87	46	60	92		45	94	79		88		87	
16								49	43	23	32	66			41	100					53	32					25						
17	96	99	99	97	97	41		25	63	99	72	94		81		22	100	73		56	62	71		77		65		54	65	95	97		
18	29	90	73	32	67	74		68	38	93	28	73			21		75	100		22		21		75			29		93	31	58		55
19		42		22					43	86	27	49		71	23		52		100		68										22		
20	37	79	23	87	38		21	40	27	88		27					55			100											22		
21	22	96	80	92	84	22	73	38	97	96	78	87		51	90	33	56		39	21	100	73		30		38		91		56	69		
22		79	77	75	59			79	69	91		36		76	48	31	74				79	100	29			65		49		56			
23	25	93	37	75	43			41	69	92		67		51	59							21	100			30		33		23			
24	43	96	95	94	87			65	82	92		60			92		76	70			39	24		100			67	85		87	76		67
25		21																							100								
26		90		84	83			49	34	89	36	92		72	45		63				36	63	28			100	29			50			
27	79	94	34	91	73		23	85	57	93	49	73		29	93			22						60			100	92			70		78
28	75	95	52	81	75			93	88	97	54	66			71		57				93	50	30	80			93	100	29	39	49		
29	95	96	95	94	94			24	22	94	65	37					73	94										53	100		94		
30	96	99	98	96	94			82	55	97	66	90		90	90		95	31			65	67	50	85		31		51		100	92		
31	100	99	99	99	98		78		60	99	97	94		72			97	64			67			81			71	41	91	89	100		62
32											32				81												29					100	
33	36	73	65	67	82		22	92	32	73	43	22						56						65			78				65		100

Fig. 6: Pairwise profile-profile sequence comparison of $(\beta\alpha)_8$-barrel superfamilies. The query superfamilies are listed at the edges using the specific superfamily identifiers as labels. The label prefix: c.1 (class α/β and $(\beta\alpha)_8$-barrel -barrel) was omitted for clarity. Each column indicates which other $(\beta\alpha)_8$-barrel fold superfamily from the dataset is detected by HMM-HMM in a probability range of 100% to 20% (100-80 dark red highlighted).

Let’s look at the connectivity starting from the weakest linked superfamily: the Monomethylamine methyltransferase MtmB superfamily (SCOP c.1.25) that only hit with 20% probability the Ribulose-phoshate binding barrel superfamily (SCOP c.1.2). The connection is not bidirectional, which means that no HMMs representing the Ribulose-phosphate binding barrels hit with more than 20% probability any member of the Monomethylamine methyltransferase MtmB superfamily. TM1631-like (SCOP c.1.32), NAD(P)-linked oxidoreductase (SCOP c.1.7), Malate synthase G (SCOP c.1.13), Bacterial luciferase-like, (SCOP c.1.16), cobalamin (vitamin B12)-dependent enzymes (SCOP c.1.19), tRNA-guanine transglycosylase (SCOP c.1.20) are weakly connected. The rest of the superfamilies were connected multiple times, with the phosphate binding superfamilies being the most highly connected cluster (SCOP c.1.1-6).

Hence, our data supports a monophyletic origin for most $(\beta\alpha)_8$-barrel superfamilies. Our results were consistent with previous reports. The systematic comparison of the data was not trivial. Earlier studies were based on a combination of diverse approaches (PSI-BLAST, structural and functional comparisons) in order to evaluate the implication of the detected similarities among the $(\beta\alpha)_8$-barrel superfamilies. In contrast, we only employed HMM-HMM comparisons at superfamily level. Second, there are major differences between the database we used in our work (SCOP) and the database used by Nagano and coworkers (CATH). An illustrative example of this problem is constituted by the structure of the catalytic domain of a thermophilic endocellulase (PDB code 1TML); this structure is classified in CATH as $(\beta\alpha)_8$-barrel, while in SCOP it is a 7-stranded β/α barrel. Finally, there are novel $(\beta\alpha)_8$-barrel structures in PDB database and these proteins are classified in new superfamilies.

The work from (*31*) is the most inclusive study of this type; therefore, we compared our results with their findings. To overcome the problem of using different classification schemes, we performed a cross association using the PDB identifiers listed in Nagano’s research, see table 9 for details. Our work fully recreates the previously observed connections. Overall in our study 30 of the 33 $(\beta\alpha)_8$-barrel superfamilies are connected with more than 80% bidirectional probability, (figure 6).

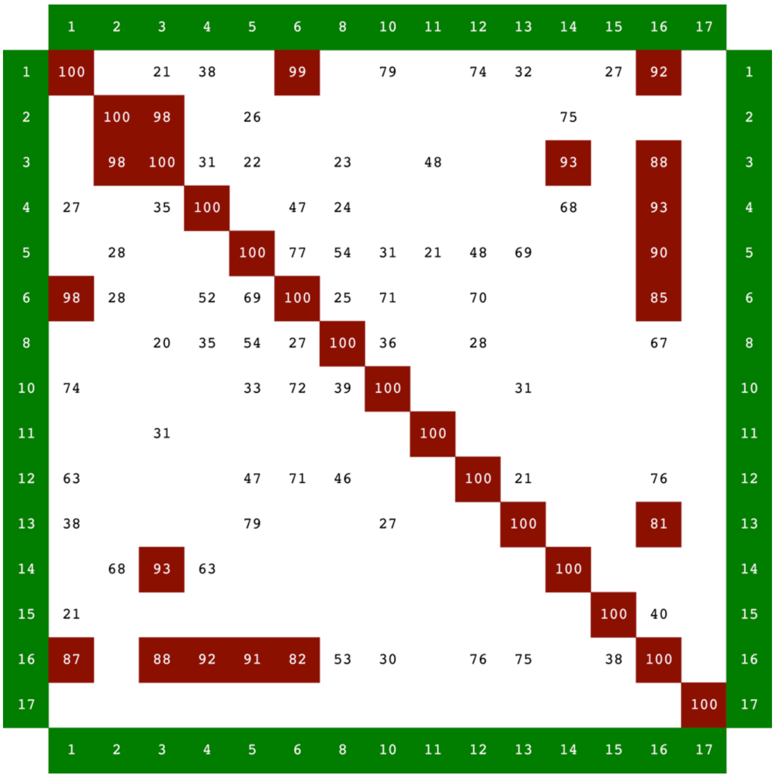
The data shown in our study strongly suggests a common origin for 30 $(\beta\alpha)_8$-barrel superfamilies. Interestingly, the Aldolase superfamily (c.1.10) is the most highly connected hub: it hits 25 of the 33 superfamilies with a probability better than 80 %. Moreover, the class I aldolase family is proposed in SCOP as a possible link between the aldolase superfamily and the phosphate-binding $(\beta\alpha)_8$-barrels. The protein structures of the Cyclase domain of the imidazoleglycerolphosphate synthase HisF and the Phosphoribosylformimino-5-aminoimidazole carboxamide ribotite isomerase HisA are classified into this superfamily. These protein structures are the most cited examples for the $(\beta\alpha)_8$-barrel evolution via duplication and fusion of a $(\beta/\alpha)_4$ unit.

Table 9: Comparison of earlier studies and the current research

Superfamily as defined in (*31*)	PDB identifier	Previous work	SCOP & ASTRAL identifier	Present work
Alanine racemase (ALR)	1BD0	Connected	c.1.6.1 d1bd0a2	Connected
Dihydropteroate (DHP) synthetase (DHPS)	1AD4	Connected	c.1.21.1 d1ad4b_	Connected
FMN dependent fluorescent proteins	1FVP	Connected	c.1.16.2 d1fvpa_	Connected
Luciferase-like proteins LUCL	1LUC 1LUC	Connected	c.1.16.1 d1lucb_ c.1.16.1 d1luca_	Connected
Seven-stranded glycosidases 7CEL	1TML 1CB2	**Not connected**	c.6.1.1 d1tmla_ c.6.1.1 d1cb2a_	Connected
Phoshoenolpyruvate (PEP) binding enzymes (PEPE)	1PKM 1DIK0	Connected	c.1.12.1 d1pkma2 c.1.12.2 d1dika	Connected
Aldolase class I family (ALD1)	1NAL1 1DHP 1FBA	Connected	c.1.10.1 d1nal1_ c.1.10.1 d1dhpa_ c.1.10.1 d1fbaa_	Connected
Glycosidases (GLYC) 7-1 a-Amylase (AAMY) **7-2 Endoglucanase (EG)** 7-3 Chitinase (CHTN) **7-4 Chitobiase (CHOB)**	1AVA 1BAG 1UOK 1CGT 1BYB 1CEO 1XYZ 1BQC 1CTN 2HVM 1QBA	Connected **Not connected** Connected **Not connected**	c.1.8.1 d1avaa2 c.1.8.1 d1baga2 c.1.8.1 d1uoka2 c.1.8.1 d1cgta4 c.1.8.1 d1byba_ c.1.8.3 d1xyza_ c.1.8.5 d1ctna2 c.1.8.5 d2hvma_ c.1.8.6 d1qbaa3	Connected Connected Connected Connected Connected Connected Connected Connected
Triose phosphate isomerase (TIM)	1TPF	Connected	c.1.1.1 d1tpfa_	Connected
NADP-dependent oxidoreductase (NADO)	1ADS 1LWI 2ALR	**Not connected**	c.1.7.1 d1adsa_ c.1.7.1 d1lwia_ c.1.7.1 d2alra_	Connected Connected Connected
tRNA-guanine (tRNA-G)	1WKF	Connected	c.1.20.1 d1wkfa_	Connected
Rubisco (RUB)	1RBL	Connected	c.1.14.1 d1rbla1	Connected
Enolase superfamily (ENOL)	1ONE 1MDL	Connected Connected	c.1.11.1 d1onea1 c.1.11.2 d1mdla1	Connected Connected
FMN-dependent oxidoreductase and phosphate (PP) binding enzymes (FMOP)	1FCB 1GOX 2TMD 1OYA 1AK5 2TPS 1UBS 1PII 1NSJ	Connected Connected Connected Connected	c.1.4.1 d1fcba1 c.1.4.1 d1goxa_ c.1.5.1 d1ak5a1 c.1.3.1 d2tpsa_ c.1.2.4 d1piia1 c.1.2.4 d1ubsa_	Connected Connected Connected Connected
Metal-dependent hydrolases (MHYD)	1A4M 2KAU 1PSC 1BF6	Connected Connected	c.1.9.1 d1a4ma_ c.1.9.2 d2kauc2 c.1.9.3 d1psca_ c.1.9.3 d1bf6a_	Connected Connected
Divalent-metal-dependent enzymes (xylose isomerase-like Proteins) (XYLL)	1XIB 1A0D 1QTW	Connected Connected	c.1.15.3 d1xiba_ c.1.15.3 d1a0da_ c.1.15.1 d1qtwa_	Connected Connected
Aldolase class II (ALD2)	1B57	Connected	c.1.10.2 d1b57a_	Connected
Phosphatidylinositol (PI) phospholipase C (PIPLC)	1GYM 1QAS 2PLC	**Not connected** **Not connected** **Not connected**	c.1.18.2 d1gyma_ c.1.18.1 d1qasa3 c.1.18.2	Connected Connected Connected
Quinolinic acid phosphoribosyl(QAPR) transferase (QAPRT)	1QPO	Connected	c.1.17.1 d1qpoa1	Connected

3.2 Flavodoxin-like fold evolutionary relationships

Our analysis is the first study that explores whether the superfamilies that adopt the flavodoxin-like fold share a common evolutionary origin. In contrast to the observed superfamily connectivity pattern of the $(\beta\alpha)_8$-barrel fold, the graph in figure 7 shows few high probability bidirectional connections. The Precorrin-8X methylmutase CbiC/CobH superfamily (SCOP c.23.17) did not hit any other superfamily with more than 20% probability. In fact, only 8 of 15 superfamilies were connected by 80% bidirectional probability hits. The Class I glutamine amidotransferase-like superfamily is a hub that connects most of the superfamilies.



	1	2	3	4	5	6	8	10	11	12	13	14	15	16	17	
1	100		21	38		99		79		74	32		27	92		1
2		100	98		26							75				2
3		98	100	31	22		23		48			93		88		3
4	27		35	100		47	24					68		93		4
5		28			100	77	54	31	21	48	69			90		5
6	98	28		52	69	100	25	71		70				85		6
8			20	35	54	27	100	36		28				67		8
10	74				33	72	39	100			31					10
11			31						100							11
12	63				47	71	46			100	21			76		12
13	38				79			27			100			81		13
14		68	93	63								100				14
15	21												100	40		15
16	87		88	92	91	82	53	30		76	75		38	100		16
17															100	17
	1	2	3	4	5	6	8	10	11	12	13	14	15	16	17	

Fig. 7: Pairwise profile-profile sequence comparison of flavodoxin-like fold superfamilies. The query superfamilies are listed at the edges using the specific superfamily identifiers as labels. The label prefix: c.23 (class α/β and flavodoxin-like fold) was omitted for clarity. Each column indicates which other flavodoxin-like fold superfamilies from the dataset are detected by HMM-HMM in a probability range of 100% to 20% (100-80 dark red highlighted).

In contrast to what has been shown regarding the evolutionary origin of the $(\beta\alpha)_8$-barrel fold, little is known about the evolutionary scenario behind the flavodoxin-like fold superfamilies. We discovered a strongly supported sequence-based connection that indicates a homologous relationship between six flavodoxin-like fold superfamilies using profile-profile comparisons.

The enzymes that belong to the Class I glutamine amidotransferase-like superfamily (SCOP c.23.16) remove the ammonia functional group from glutamate to transfer it into a specific substrate. This superfamily reveals itself as a hub connecting five other flavodoxin-like fold superfamilies, with more than 80% HHsearch bidirectional probability: CheY-like (SCOP c.23.1), Hypothetical protein MTH538 (SCOP c.23.3), succinyl-CoA synthetase domains (SCOP c.23.4), flavoproteins (SCOP c.23.5) and the B12-binding domain superfamily (SCOP c.23.6).

Although the Class I glutamine amidotransferase-like superfamily was the most connected group of proteins, we will discuss in detail the sequence based connection between the response regulators (SCOP c.23.1) and the B12-binding domains (SCOP c.23.6) because these superfamilies were strongly related with $(\beta\alpha)_8$-barrel fold superfamilies, as will be discussed in the next section.

CheY (one of the most studied member of the c.23.1 superfamily) is activated by auto-phosphorylation at a conserved aspartic acid after the recognition and binding to the P2 domain of CheA. Later on, it interacts with its target FliM to trigger flagella movement. Finally, it becomes dephosphorylated by CheZ. In contrast, the B12-binding domains (SCOP c.23.6) are found in proteins that perform distinct kinds of reactions at the cobalt carbon bond. These proteins bind B12 using a conserved histidine in a DXHXXG sequence. Thus, both flavodoxin-like superfamilies carry out very different activities although they belong to the same fold and appear to be related.

A profile-profile alignment with 113 aligned columns and 96.93 HHsearch probability between CheY (PDB 2PMC, from *Salmonella enterica subsp. enterica serovar Typhimurium*) and the B12-binding domain (PDB 7REQ, Methylmalonyl-CoA Mutase from *Propionibacterium freudenreichii subsp. shermanii*) revealed 10 identical.

As expected, given that both proteins have very different associated functions, no apparent conserved functional residues are revealed by the profile-profile alignment. Unexpectedly, the first beta strand is not matched in this sequence-based alignment; in contrast, in a pure structure-based alignment of the two proteins, the first beta strand is matched (97 C-alpha carbons were structural superimposed with an RMSD of 2.3). This discrepancy between sequence and structure-based findings suggests that β1 is a late or divergent embellishment of the fold.

The superposition shows a conserved proline (P706 in 7req and P110 in 2mpc) that seems to be important in order to build a turn from β5 to α5 in both proteins. The structure-based alignment also reveals that the β/α loops are longer in the B12-binding domain than in the response regulator, probably due to functional reasons. 7req interacts with cobalamin via key residues that are present in the above-mentioned loops.

A leucine-glycine-phenylalanine motif conserved in both proteins (located in the α1/β2 loop) seems to play a structural role, since no functional residues have been mapped there. The highly conserved histidine (that binds the cobalt of the B12 molecule) from the B12-binding domain is in an equivalent position of a phenylalanine in the response regulator. Moreover, the phosphorylation site in CheY (ASP57) is replaced by a serine (that establishes a hydrogen bond with the nucleotide part of the B12 molecule) in the B12-binding domain.

The conserved glycine, which establishes hydrophobic interactions with the nucleotide part of the B12-binding domain, was aligned with a threonine in its equivalent position on the response regulatory protein. In contrast to the evolutionary scenario drawn in the barrel superfamilies, where many functional features that include a phosphate-binding site are linking different superfamilies, these two flavodoxin-like fold superfamilies show an evolutionary signal present only in the global profile-profile alignment. This outcome is expected, given that both superfamilies are performing rather different functional roles.

One possible explanation for the observed sequence based similarities without any apparent functional connection, is that both superfamilies evolved to be specialized and fulfill two different functions. From this perspective, the comparison performed with these two superfamilies offers a simple method to discriminate between functional and structural important residues in protein scaffolds.

3.3 Sequence-based evidences of homology between $(\beta\alpha)_8$-barrel and flavodoxin-like fold superfamilies.

Having explored the intra-fold evolutionary relationships we then recorded the hits between the two folds. Figure 8 shows the results of searching the SCOP95 database with HMM profiles of all $(\beta\alpha)_8$-barrel (left) and flavodoxin-like fold (right) structures. We plotted density of the HHsearch hits versus probability (see figure 8 legend for details). Remarkably, when we launched a search starting with a $(\beta\alpha)_8$-barrel query, at around 75 % probability the number of flavodoxin-like fold hits is comparable with the number of $(\beta\alpha)_8$-barrel fold hits (self-hits). The opposite case occurs around a 90 % probability cut-off. What is more, searches with flavodoxin-like fold queries at 75 % probability cut-off hit as much as two times more $(\beta\alpha)_8$-barrels than any other fold. A heat-map of the maximum probabilities reached between $(\beta\alpha)_8$-barrel and flavodoxin-like superfamilies is provided in figure 9. Lines are labeled with the SCOP identifiers and the legend indicates the color scale associated with a probability score.

We analyzed in detail the inter-fold hits to find superfamilies from the different folds that share a common evolutionary origin. We discovered that only two flavodoxin-like fold superfamilies hit with more than 80% bidirectional probability a number of $(\beta\alpha)_8$-barrel superfamilies: the CheY-like (SCOP c.23.1) and the B12-binding domains (SCOP c.23.6) (see figures 9 and 10 for details).

The observation is remarkable because the flavodoxin-like fold is populated by 15 homologous superfamilies, and only two of them showed connections with $(\beta\alpha)_8$-barrel superfamilies. Therefore we can argue that the high probability hits between different folds is not only due to the high structural similarity of the flavodoxin-like fold with the $(\beta\alpha)_8$-barrel fold. For instance, the CheY-like (SCOP c.23.1) superfamily hit with 93% (HHsearch probability) the ribulose-phoshate binding barrels (SCOP c.1.2). The probability with which the ribulose-phoshate binding barrel hit the CheY-like superfamily was 88% (see figures 9 and 10 for details). This data is very suggestive evidence for common ancestry of these two superfamilies. The CheY-like superfamily hit with 94% HHsearch probability the ThiG-like superfamily. These enzymes are involved in the biosynthesis of thiamine and share the same phosphate binding site with other $(\beta\alpha)_8$-barrel superfamilies.

The ThiG-like superfamily (SCOP c.1.31) hit the CheY-like (SCOP c.32.1) superfamily with >90% HHsearch probability. And the B12-binding domain (SCOP c.23.6) also hit with very high probabilities many $(\beta\alpha)_8$-barrel fold superfamilies; it was aligned with high (>80%) bidirectional probability to 12 $(\beta\alpha)_8$-barrel fold superfamilies. Some of the $(\beta\alpha)_8$-barrel fold superfamilies that are strongly related with the B12-binding domains are: the phosphate binding barrels, aldolases, the phosphoenolpyruvate/pyruvate domains, nicotinate/quinolinate PRTase C-terminal domain-like, and ThiG-like among others.

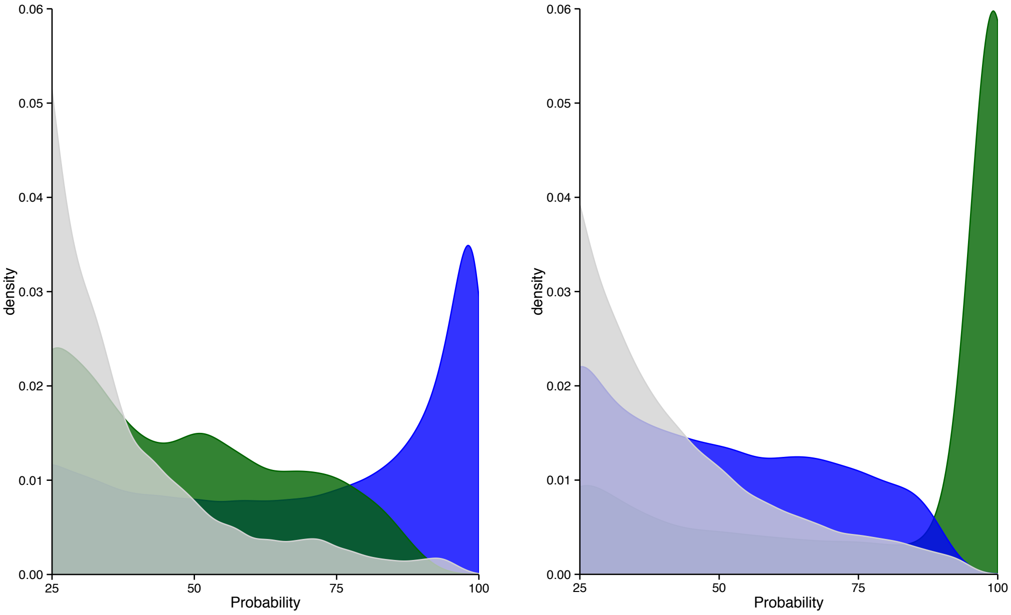


Fig. 8: The SCOP95 database represented by HMM profiles was searched with the HMM profiles of all $(\beta\alpha)_8$-barrel and flavodoxin-like fold structures. We plotted density of the HHsearch hits versus probability. Left: hits of $(\beta\alpha)_8$-barrel to other $(\beta\alpha)_8$-barrel (blue), to flavodoxin-like (green) and to all other folds (grey). Right: hits of flavodoxin-like to other flavodoxin-like (green), to $(\beta\alpha)_8$-barrel (blue) and to all other folds (grey).

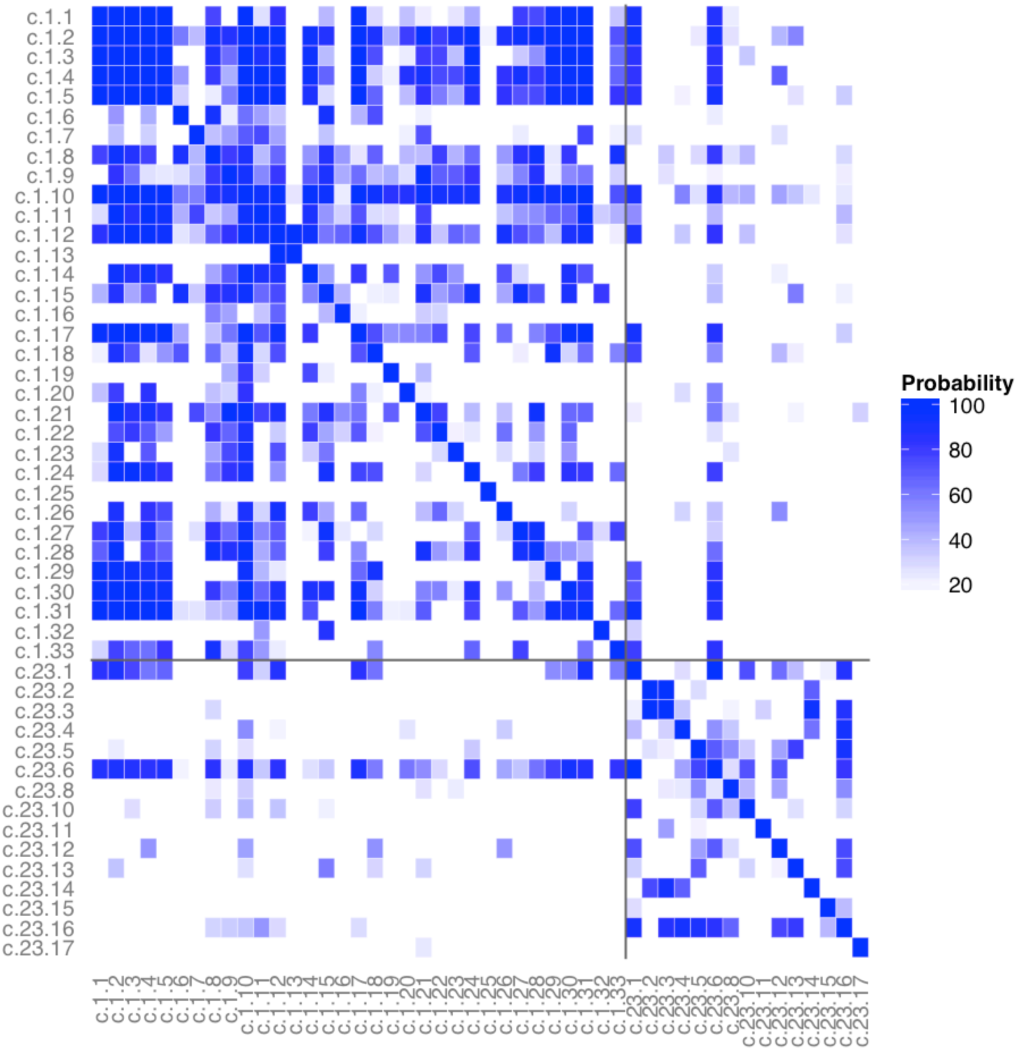


Fig. 9: Heat-map of the maximum probabilities reached between (βα)$_8$-barrel (c.1) and flavodoxin-like (c.23) superfamilies. Lines are labelled with the SCOP identifiers and the legend indicates the colour scale associated with a probability score.

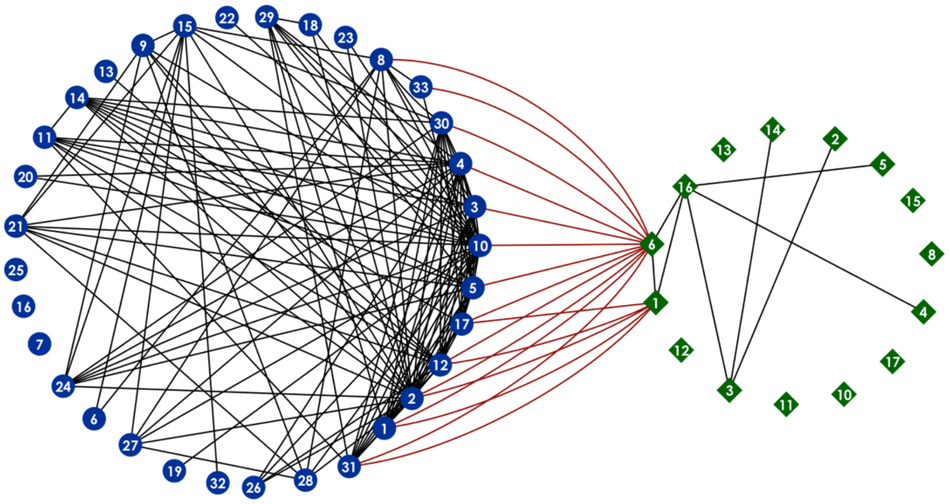


Fig. 10: Sequence-based exploration of all SCOP homologous groups that fold into the $(\beta\alpha)_8$-barrel (33 homologous superfamilies represented by blue circles) and flavodoxin-like folds (15 superfamilies represented by green diamonds). The black (intrafold hits) and red (interfold hits) edges are bidirectional HHsearch hits with probabilities higher than 80%.

In figure 9 we can see many more hits between flavodoxin-like and $(\beta\alpha)_8$-barrel fold superfamilies. Most of these hits were recorded with low probability, being clearly weaker than the connections established by the CheY and B12-binding superfamilies. Our analysis allowed us to identify which superfamilies from different folds are more closely related. The representation shown in figure 10 clearly demonstrates the strong sequence based relationship between the superfamilies form different folds. Together, our findings represent the first indication of common ancestry for these folds. Also show which homologous groups are located at the sequence space interface of the $(\beta\alpha)_8$-barrel and the flavodoxin-like fold. After finding that the two super-folds are related, we wanted to know how one fold could have converted into the other. Therefore we launched remote homologous searches to explore the vast sequence space that separates these two ancestral folds.

3.4 Searching for sequences with intermediate features between two super folds.

Once we found sequence-based evidences for homology between the two super-folds, we wanted to know how one fold could have interconverted into the other. In order to explore this possibility, we searched the huge sequence space yielded by the recent mass sequencing genome projects.

Folds can be populated by certain sequences that are compatible with them. Protein sequences for which there are no structural representatives can be compared with homologous sequences that have structures deposited in the protein data bank (PDB). Generally speaking, the structure or function of a novel sequence protein can be deduced based on homology to known proteins. A protein sequence with unknown structure, which is 35% identical to another homolog that folds as a β-barrel, is very likely to be folded as a β-barrel as well. We theorized that by searching for ambiguous sequences, specifically those sequences that are equally related to members of a $(\beta\alpha)_8$-barrel and a flavodoxin-like structures, we would find interesting information on the evolution of these folds.

It is very unlikely to find amino acid sequences that are more than 30% identical, over the entire polypeptide chain, to two folds simultaneously. Only some special cases are documented: two were engineering experiments (*65, 66*) to specifically explore this possibility and one more was identified by a stepping stone method (*67*). Our approach is fundamentally different, since we want to systematically find natural sequences that can be seen as the fossil record of fold evolution.

If we start a BLAST search with a protein sequence for which the structure is known to be a $(\beta\alpha)_8$-barrel we can be confident that most of the homologous hits, up to certain identity threshold, would also adopt a $(\beta\alpha)_8$-barrel fold. For instance, if a pairwise sequence alignment between protein X and Y, for which only Y has a known 3D structure, shows that both proteins share 35% of identical residues we can then imply that protein X has the same fold as Y. The key point of our research is to find sequences that are related, but not too closely, with sequences that fold into $(\beta\alpha)_8$-barrel and flavodoxin-like fold structures.

If we imagine the sequence space as an ocean, we have to sail the sequence space away from the $(\beta\alpha)_8$-barrel structural island towards the flavodoxin-like fold structural island. We wanted to solve the structure of the sequences that were located in the middle of our journey in order to gain structural insight on the fold change path between both folds.

Once we analyzed the profile-profile pairwise comparisons we launched HHsenser searches using few sequences from the $(\beta\alpha)_8$-barrel and flavodoxin-like fold structures that are most closely related. For further work, it will be necessary to launch searches not only with one seed per superfamily but also many different starting sequences. We recognize that this is a necessary analysis, because the sequences from the structures in the same superfamily are highly diverse.

For instance, the sequences from 1I9C, 7REQ, 1BMT and 2YXB (all protein structures classified into the c.23.6 superfamily, or B12-binding domains) only cluster together if the sequence identity among all the proteins in the cluster goes as low as 25%. In other words, 1I9C only shares 25% identical residues with 1BMT. For further work it is necessary to launch a BLAST search using the sequence from one B12-binding domain structure, organize the hits by sequence identity and launch HHsenser searches using all non-redundant hits (e.g. 1I9C and 1CCW are 98% identical, one has the ligand bound). By following this procedure the possibility of missing informative intermediate hits is reduced.

I am going to explain in detail the selection procedure of some targets, illustrating the difficulty to outline a general systematic methodology for choosing targets. I first started by collecting several HHsenser outputs produced by using $(\beta\alpha)_8$-barrel and flavodoxin-like fold sequences as queries. By launching HHsenser, you can have access to different kinds of alignments, I mainly used two: a large permissive alignment (where many homologous hits could be found, and sometimes non homologous spurious hits are included) and a reduced version of the previous alignment (including the 100 most diverse hits).

I used this last type of reduced alignment to build a clustering map in which two clusters of flavodoxin-like fold superfamilies (c.23.6 and c.23.1) are located close to the core of $(\beta\alpha)_8$-barrel superfamilies. This data is not included due to the redundancy of the experiment.

For searching the intermediate candidates I set very relaxed parameters of the HHsenser search engine to include as many homologous hits as possible (see methods for details). The sequences listed in table 10 were used as queries to launch HHsenser searches.

Table 10: Selected queries to perform HHsenser searches of remote homologous sequences. It is listed the superfamily in which the query sequences are classified.

$(\beta\alpha)_8$-barrel and Flavodoxin-like fold superfamilies	Superfamily name (selected sequences)
c.1.2	Ribulose-phoshate binding barrel (hisF, hisA, IGPS)
c.1.3	Thiamin phosphate synthase (thiamine synthase)
c.1.5	Inosine monophosphate dehydrogenase (IMPDH)
c.1.33	EAL domain-like
c.1.17	Nicotinate/Quinolinate PRTase C-terminal domain-like
c.23.1	CheY-like
c.23.6	B12-binding domain

We searched the non-redundant and the environmental sequences databases using the queries indicated in table 10. HHsenser can detect thousands of homologous proteins. Interestingly, the searches using the sequence from CheY (*Thermotoga maritima)* produced around 22,000 hits.

This is a special case of sequence-function relationship, since an infrequently high number of non-redundant sequences are compatible with a single fold that shows a narrow well defined specific function (flavodoxin-like fold, response regulator). So far, the sequence from CheY was the query that produced the most hits of all the queries tried.

We collected two HHsenser permissive alignments, one produced by a $(\beta\alpha)_8$-barrel fold query (hisF; SCOP c.1.2, for instance) and another alignment produced by a flavodoxin-like fold query (CheY ; SCOP c.23.1, for instance). We merged them into a single input file to feed the clustering program CLANS.

CLANS is a program implemented as a variant of the *Fruchterman and Reingold* graph layout algorithm (*35*) that provides representations of pairwise sequence similarities. Dots represent sequences in a three-dimensional map; high scoring segment pairs (from BLAST/PSI-BLAST) are shown as edges that connect the dots. Attractive forces are proportional to the negative logarithm of the high scoring segment pairs (HSPs) P-values. The lower the P-value the stronger the attractive force. Between 3000 and 9,500 sequences composed the regular input files for CLANS.

The upper limit of clustering in a desktop computer was 10,000 sequences; we clustered up to 9500. Using more than this number caused the graphical interface of CLANS to freeze. The program performs BLAST searches first; subsequently, using a graphical interface, the user starts the actual clustering.

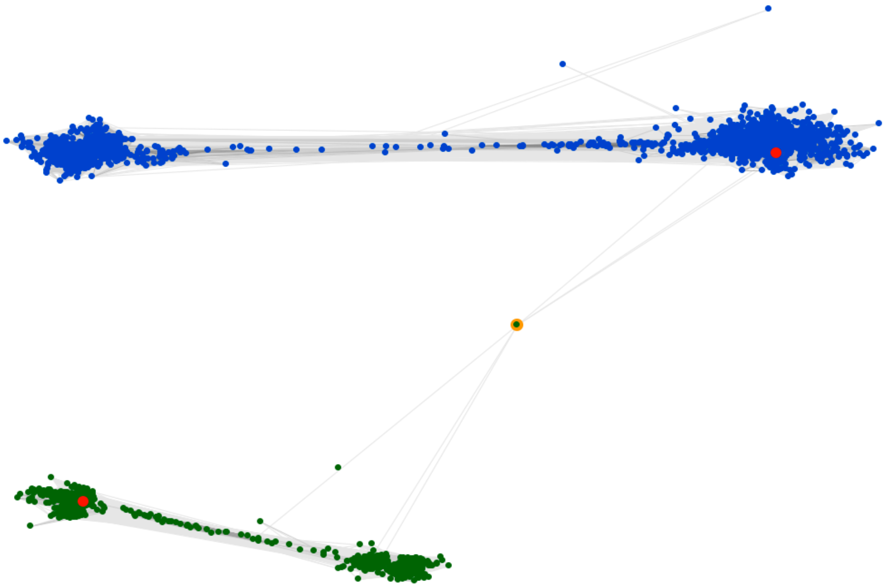


Fig. 11: CLANS clustering map showing flavodoxin-like fold hits in blue, and $(\beta\alpha)_8$-barrel hits in green. The starting queries are highlighted as red circles. The sequence (sub-domain size fragment, 31-163 from 165 residues) of the glutamate mutase S-chain *from Streptomyces halstedii* is highlighted in orange/green right at the middle of the cluster map. The high scoring pair hits from the BLAST searches, at a better P-value of 1.0e-8, are shown as connecting edges.

The clustering map in figure 11 shows flavodoxin-like fold hits colored in green, and $(\beta\alpha)_8$-barrel hits colored in blue. The flavodoxin-like fold query was the sequence from 2YXB, while the $(\beta\alpha)_8$-barrel query was 1THF. It shows HSPs represented as edges that connect the dots at a P-value better than 1.0e-08. As expected, 1172 flavodoxin-like hits (green) cluster together; it is the same case for 3040 $(\beta\alpha)_8$-barrel hits (blue). The hits were found in the nr and metagenomic sequences databases.. More interestingly, the clusters from different colors seem spliced into sub-clusters.

The configuration of the clusters reflects the deepness of the remote homologous search performed with HHsenser. It shows how a search started with Hisf from *thermotoga maritima* (highlighted as the red dot confined in the blue cluster at the upper right), can reach the homologous family HisA (blue cluster, upper left). In the case of the flavodoxin-like clusters (green) the search started with the B12-binding domain (highlighted as the red dot confined in the green cluster at bottom left), of the methylmalonyl-CoA mutase alpha-subunit from *Aeropyrum pernix,* reached B12-binding domains from the methionine synthase family (green cluster at the bottom center).

The sequence that connects clusters from different colors, highlighted as a green/orange dot, is a fragment (131 residues from 165) of the glutamate mutase S-chain from *Streptomyces halstedii*. As the first step performed to evaluate whether this protein could be considered as an intermediate hit, we searched the PDB database (using BLAST) with the short fragment. It showed sequence similarities to flavodoxin-like fold structures (2XYB, 2GKG) and $(\beta\alpha)_8$-barrel structures, see table 11 for details.

Table 11: Hits of a BLAST search launched using the sequence fragment (131 residues from 165) of the glutamate mutase S-chain from *Streptomyces halstedii*. The identities over the aligned region are listed, and the E-values of the hits are also shown.

Hits (number in hit-list)	Identities	E-value
1. 2YXB, Coenzyme B12-dependent mutase SCOP c.23.6	25/99 (25%)	Expect = 3e-04
2. 2GKG, Response regulator homolog SCOP c.23.1	24/78 (30%)	Expect = 0.001
8. 1VRD, IMPDH SCOP c.1.5	20/66 (30%)	Expect = 0.081

By resubmitting the search with the complete sequence, we found a different distribution of values for the hits; moreover, a different $(\beta\alpha)_8$-barrel structure is found among the top hits (table 12).

Table 12: Hits of a BLAST search launched using the whole sequence (165 residues) of the glutamate mutase S-chain from *Streptomyces halstedii*. The identities over the aligned region are listed, and the E-values of the hits are also shown.

Hit (number)	Identities	E-value
1. 1CCW, Protein (glutamate mutase) SCOP c.23.6	29/122 (23%)	Expect = 1e-05
2. 2GKG, Response regulator homolog SCOP c.23.1	19/47 (40%)	Expect = 2e-04
11. 2HTM, Thiazole biosynthesis protein SCOP c.1.31	30/101 (29%)	Expect = 0.89

The size of the intermediate query used to perform a back validation search is going to strongly influence the list of hits. The first hits shown in tables 11 and 12 suggest that this protein may adopt the flavodoxin-like fold; nonetheless, there are also high local similarities to $(\beta\alpha)_8$-barrel fold structures.

Given the size and BLAST result, we could assume that this novel protein sequence will adopt the flavodoxin-like fold. However, it also shows some local sequence similarities to $(\beta\alpha)_8$-barrel structures. We then performed a secondary structure prediction (Quick 2D server MPI toolkit) on this intermediate sequence. The server predicted a mixed beta/alpha protein with the following topology: the first 20 residues are predicted as disordered and then we found four beta/alpha modules $(\beta\alpha)_4$.

The glutamate mutase S-chain from *Streptomyces halstedii* is therefore chosen for experimental characterization because: a) it shows similarities with the two folds analyzed in this work; b) despite the similarity in size to the flavodoxin-like fold, the secondary structure prediction of the protein $[(\beta\alpha)_4]$ suggests a possible dimeric arrangement that may fold in some way similar to a $(\beta\alpha)_8$-barrel.

The previously discussed remote homologous search and clustering process, was not deep enough to allow an effective search of sequences with novel structural features. Therefore, we selected the sequences from the glutamate mutase from *Clostridium cochlearium* (SCOP c.23.6; PDB 1I9C) and the inosine monophosphate dehydrogenase from *Streptococcus pyogenes* SCOP c.1.5; PDB 1ZFJ) to perform a new search of the non-redundant and metagenomic sequences databases.

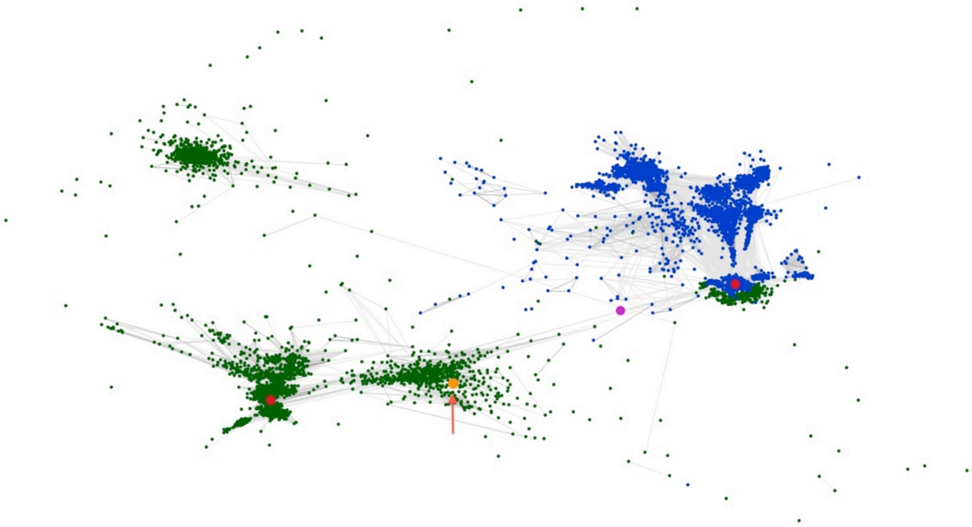


Fig. 12: CLANS clustering map showing 4094 flavodoxin-like hits colored in green and 5116 $(\beta\alpha)_8$-barrel hits colored in blue. The starting queries are highlighted as red circles. The high scoring pair hits from the BLAST searches, at a better P-value of 1.0e-10, are shown as connecting edges. The orange circle (emphasized by an arrow) is the sequence of the hypothetical protein TM0182 from *Thermotoga maritima*, while the pink circle is the sequence of the hypothetical protein STH347 from *Symbiobacterium thermophilum*

The clustering map in figure 12 shows 4094 flavodoxin-like hits colored in green and 5116 $(\beta\alpha)_8$-barrel hits colored in blue. The sequences were compared by BLAST and 5000 rounds of clustering were performed; the queries for the HHsenser search are highlighted as red circles. There are two more sequences highlighted as circles: the orange circle is the sequence of the hypothetical protein TM0182 from *Thermotoga maritima*, while the pink circle is the sequence of the hypothetical protein STH347 from *Symbiobacterium thermophilum*. BLAST hits (better 1.0e-10 P-value) are shown as edges in the map.

The first interesting feature to analyze in the CLANS map in figure 12 is the overlapping of two clusters from different colors around the $(\beta\alpha)_8$-barrel query. Remarkably, the search launched with a flavodoxin-like fold query reached the $(\beta\alpha)_8$-barrel sequence space: sequences annotated as inosine-5'-monophosphate dehydrogenases (IMPDH) were hit; the IMPDH function is associated with the $(\beta\alpha)_8$-barrel fold.

The configuration of this cluster reveals that one of the interfaces between both folds is indeed located between the B12-binding (flavodoxin-like fold c.23.6) and the IMPDH [$(\beta\alpha)_8$-barrel c.1.5.1] superfamily. The sequence space search, started on the flavodoxin-like fold structural island, hit many sequence intermediates to reach the $(\beta\alpha)_8$-barrel fold structural island. We performed a BLAST search using one of them: the hypothetical protein STH347 from *Symbiobacterium thermophilum*. This sequence is a 91 residues fragment of a 254 full-length protein; moreover, 17 residues at random positions are missing in this short fragment. This is a problem of the HHsenser searches, where the hits of the alignments are missing residues.

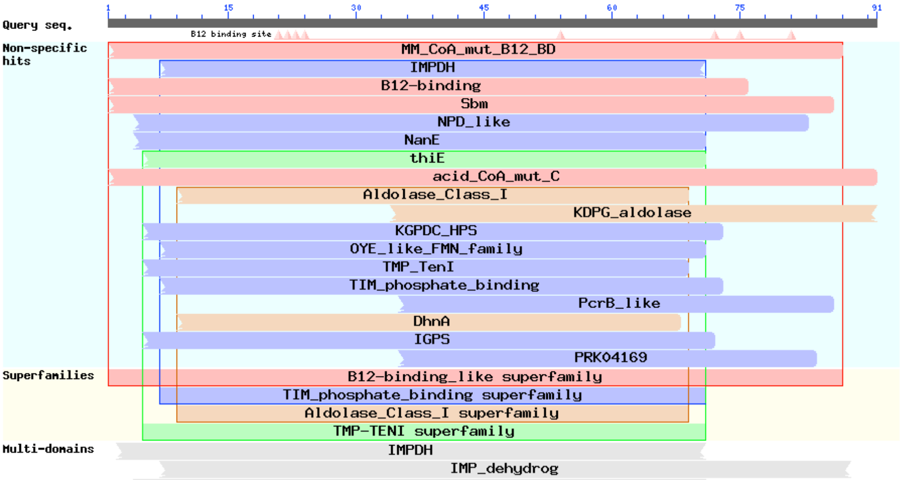


Fig. 13: BLAST result of searching the CDD using the sequence from the hypothetical protein STH347 from *Symbiobacterium thermophilum* (91 residue fragment).

The BLAST search with the short raw fragment displays a remarkable mixed signal towards flavodoxin-like and $(\beta\alpha)_8$-barrel families of proteins (figure 13). MM_CoA_mut_B12_BD, B12-binding, acid CoA_mut_C and Sbm are flavodoxin-like fold families. NanE, IGPS, thiE, aldolase and TIM_phosphate_binding are $(\beta\alpha)_8$-barrel families.

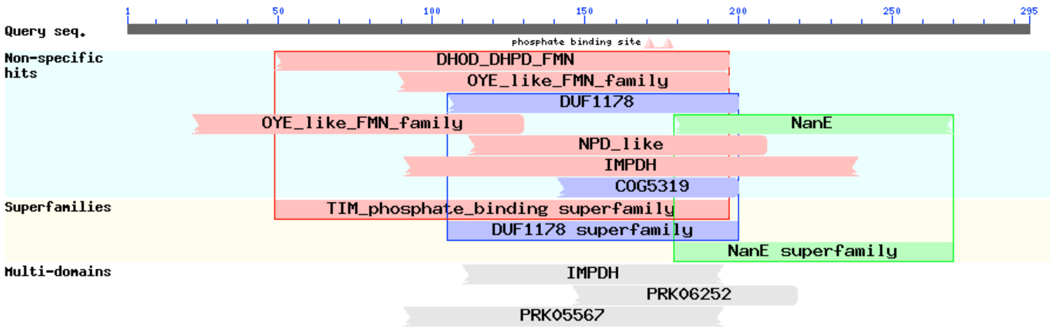


Fig. 14: BLAST result of searching the conserved domain database using the full sequence from the hypothetical protein STH347 from *Symbiobacterium thermophilum* (254 full-length sequence).

In a subsequent step, we used the complete sequence to perform a new BLAST search and the signal towards the flavodoxin-like fold families vanished (figure 14). The variation of the hits found in function of the query size indicates that the mixed signal towards both folds is local. Nonetheless, taking into account the length of the full sequence, it is likely that this protein adopts the $(\beta\alpha)_8$-barrel fold. Therefore, we did not choose this target for experimental characterization with high priority.

Besides of the cluster of mixed colors, another three flavodoxin-like fold clusters can be recognized in the cluster map from figure 12. The cluster where the flavodoxin-like fold query is confined corresponds to the B12-binding domain superfamily. Interestingly, members of the CheY-like superfamily compose the group of flavodoxin-like fold hits that appears on the left corner, away from the center of the map. The HHsenser search performed with the B12-binding domain query (1I9C, flavodoxin-like fold c.23.6) was deep enough to reach the sequence space of the homologous CheY-like superfamily.

The hypothetical protein TM0182 from *Thermotoga maritima*, highlighted as an orange circle in figure 12, is present in the green cluster located close to the middle of the map. It is not possible to find statistically significant hits by searching the PDB database using this sequence as query.

Searching the non-redundant database of proteins, with the raw short fragment, we found members of the B12-binding domain-like associated with radical SAM domain family of proteins. Two domains compose these proteins: the N-terminal domain is similar to the B12-binding domain superfamily, however it lacks the signature motif Asp-X-His-X-X-Gly that binds to cobalt. The function of these proteins is not known.

The B12-binding domain_like cluster of proteins is located in the middle of the map. As previously discussed in the introduction, the function and fold of these proteins was unknown, and they show similarities towards flavodoxin-like and $(\alpha\beta)_8$-barrel structures. Therefore, we decided to experimentally characterize the N-terminal domain of the TM0182 protein, a member of the B12-binding domain-like family of proteins, for which no structure is known so far.

The cluster map in Figure 15 shows 4749 flavodoxin hits highlighted in green and 4288 $(\beta\alpha)_8$-barrel hits colored in blue. The flavodoxin-like fold query used for the HHsenser search is the sequence of 2XYB (B12-binding domain SCOP c.23.6.1) while the $(\beta\alpha)_8$-barrel query was 1VRD (IMPDH SCOP c.1.5.1). We modified several parameters of the clustering process: a) we set an arbitrary cutoff P-value 1.0e-06 to remove many singletons after the clustering; b) we doubled the repulsion values, and c) we re-positioned the sequences to cluster again for 5000 rounds. This is why the map looks cleaner than the one shown in figure 12.

In the CLANS clustering map from figure 15 we recognize: a) B12-binding domain cluster (where the flavodoxin-like query is highlighted in red), b) the response regulators cluster (upper left), the B12-binding domain-like cluster of proteins (highlighted in green/orange), c) the sequence of the putative Fe-S oxidoreductase from *Pelobacter Carbinolicus* (fragment 21-143 from 425 residues) colored in pink and d) the cluster of the IMPDH family of proteins where the $(\beta/\alpha)_8$-barrel query (highlighted in red) is confined.

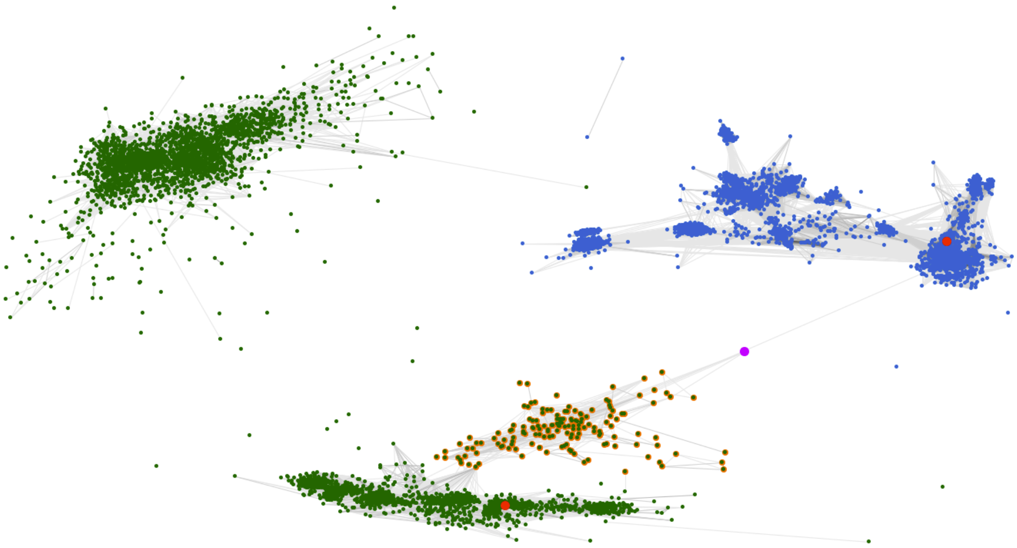


Fig. 15: CLANS clustering map showing 4749 flavodoxin hits highlighted in green and 4288 $(\beta\alpha)_8$-barrel hits colored in blue. The connections are showed at better P-value better than 1.0e-10. The queries are highlighted in red (2XYB and 1VRD). The B12-like associated with SAM radical family of proteins is highlighted in orange. The putative Fe-S oxidoreductase from *Pelobacter carbinolicus* (fragment 21-143:425) is colored in pink.

From the map it becomes clear that the sequence of the putative Fe-S oxidoreductase is an interesting target. We performed a BLAST search with the raw fragment found by HHsenser to detect a mixed signal to the B12-binding superfamily (flavodoxin-like fold associated) and to the IMPDH superfamily [(βα)8-barrel fold associated]. The putative Fe-S oxidoreductase is a multi-domain protein; the C-terminal domain is similar to SAM radical proteins while the N-terminal domain (found by the HHsenser search) shows a mixed signal towards the B12-binding superfamily and IMPDH family of proteins. The size of the full length N-terminal domain of this protein is compatible with a flavodoxin-like fold structure (150 residues); nevertheless, a $(\beta\alpha)_8$-barrel structure is found as first hit in a search of the PDB.

Remarkably, the mixed signal towards both folds was not affected by doing the BLAST search with the complete N-terminal sequence of the putative Fe-S oxidoreductase (renamed PC04) instead of using the raw small sequence found by HHsenser.

We decided to experimentally characterize this protein taking into account the following observations: a) the protein size is compatible with a flavodoxin-like fold but it hits first a $(\beta\alpha)_8$-barrel structure in a BLAST search and b) the mixed signal towards both folds is strong while doing a BLAST search with the full length domain.

As a final example of the selection procedure for the intermediate targets, we identified the hypothetical nicotinate phosphoribosyltransferase from *Aquifex aeolicus*. It was located between two clusters from different folds (see figure 16 for details). There were multiple connections directly between the two clusters from different folds; and the position of the protein was indicative of its intermediate nature. We first started doing a BLAST search with the raw short fragment identified in the CLANS clustering map. We pulled the complete sequence from the NCBI website (gi 499183116) to perform a search of the conserved domain database with the complete sequence (residues 1-426) and found similarities towards the Nicotinate phosphoribosyltransferase (NAPRTase) family of proteins. These enzymes catalyze the formation of NAMN and PPi from 5-phosphoribosy -1-pyrophosphate (PRPP) and nicotinic acid; they are present in Bacteria and Eukarya.

We searched the SCOP database to fully identify the domains present in this sequence. The N-terminal region (residues 1-131) hits with high probability the alpha/beta-hammerhead fold and the C- terminal region hits the $(\beta\alpha)_8$-barrel fold (residues 132-426). The most identical $(\beta\alpha)_8$-barrel fold structure is the putative nicotinate phosphoribosyltransferase from *Enterococcus faecalis* (SCOP c.1.17.1, PDB 2F7F). Seven parallel beta-strands compose the structure 2F7F forming an incomplete $(\beta/\alpha)_8$ barrel. The profile-profile sequence alignment between hypothetical NAPRTase from *A. aeolicus* and 2F7F extends over 291 residues, being 42% of these residues identical between both proteins. In consequence, the C-terminal domain of the hypothetical NAPRTase would likely fold into a $(\beta\alpha)_8$-barrel. Nonetheless, an HHpred search also shows a highly scored alignment between the NAPRTase and a flavodoxin-like fold structure: the methylmalonyl-CoA mutase alpha subunit, C-terminal domain from *Propionibacterium freudenreichii*, aligns over 89 residues with the NAPRTase having 29% of identical residues.

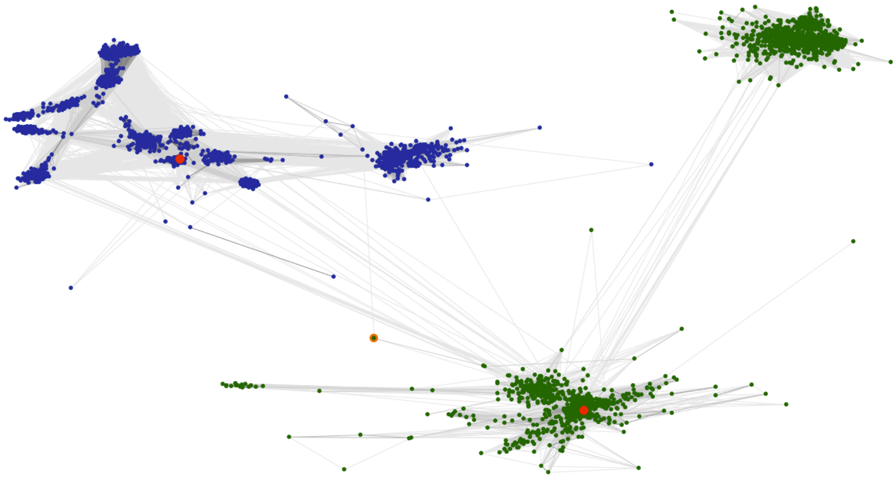


Fig. 16: CLANS clustering map showing 2425 (βα)8-barrel fold hits in blue and 1718 flavodoxin-like fold hits in green. The queries highlighted in red, were: The putative N-acetylmannosamine-6-phosphate 2-epimerase NanE from *Staphylococcus aureus* (PDB 1y0e; SCOP c.1.2.5) and the glutamate mutase, small subunit from *Clostridium cochlearium* (PDB 1CCW; SCOP c.23.6.1). The hypothetical nicotinate phosphoribosyltransferase from *Aquifex aeolicus* is highlighted in orange/green. Connections are shown a P-value of 1.0e-06.

A theoretical evolutionary path of fold change can be envisioned by using the sequence from the hypothetical nicotinate phosphoribosyltransferase (*Aquifex aeolicus*) as an anchor between the sequences of 7REQ and 2F7F. Since the pairwise alignment between the *Aquifex* sequence and 2F7F (over the entire incomplete $(\beta\alpha)_7$-barrel fold domain) is 42% identical, then is likely that the aligned region of intermediate target sequence adopts the same fold if cloned without the N-terminal domain. However this sequence is also 30% identical (locally) with the sequence of the methylmalonyl-CoA mutase alpha subunit, C-terminal domain from *Propionibacterium freudenreichii.* These 85 residues are roughly folded into an $(\beta\alpha)_3$ element corresponding to the $\alpha_5\beta_5\alpha_6\beta_6\alpha_7\beta_7$ region of 2F7F and the $\alpha_2\beta_3\alpha_3\beta_4\alpha_a\beta_5$ region of 7req. Hence this region being 42% identical to the $(\beta\alpha)_8$-barrel and 30% identical to the flavodoxin-like fold can be considered as an ancestral fragment (figure 17). This subdomain-sized fragment appears to be compatible with both folds.

It has been postulated that the $(\beta\alpha)_8$-barrel fold evolved by two fold duplication and fusion of $(\beta\alpha)_2$ elements. One could envision that the sequence signal is lost (or changed to adapt to the barrel context) in one $\alpha\beta$ element from the $(\beta\alpha)_3$ ancestral fragment. By the addition of one ab element to this ancestral $(\beta\alpha)_3$ and later duplication and fusion, a full $(\beta\alpha)_8$-barrel could be generated. On the other hand, assuming that the intermediate state between the $(\beta\alpha)_8$-barrel and the flavodoxin-like fold is a $(\beta\alpha)_4$ element, to form a flavodoxin-like fold starting from the ancestral $(\beta\alpha)_3$ element, requires major secondary structure rearrangements.

An extra sequence folded as β_1 has to invade the 5-stranded β-sheet between β_2 and β_3. Secondly, the next region of this extra sequence folded as α_1 has to flip to the other side of the 5-stranded β-sheet. Finally an extra C-terminal sequence has to form α_5 also on the same side of the 5-stranded β-sheet as α_1.

The previous evolutionary scenario is not very likely to occur. It is more likely that a sequence before α_5 in the barrel and α_1 in the flavodoxin-like fold shows high plasticity and could be folded as either β-strand or α-helix in function of the structural context. A similar scenario would be required after β_7 in the barrel and β_5 in the flavodoxin-like fold to fully recreate a flavodoxin-like fold from a sub-domain size $(\beta\alpha)_7$-barrel sequence, see figures 17 and 18 for details.

This is exactly the case of the NAPRTase from *Aquifex aeolicus*, because its secondary structure prediction seems to be ambiguous in the required regions (aligned with the full NAPRTase 2F7F). Since the 2F7F $(\beta\alpha)_8$-barrel seems to be already an intermediary step on the path towards a flavodoxin-like fold, one could imagine a scenario where cloning the region corresponding to the size of the flavodoxin-like fold, followed by either directed evolution or computational design in the ambiguous regions, this protein could be folded as a flavodoxin-like fold (see figure 18 for details).

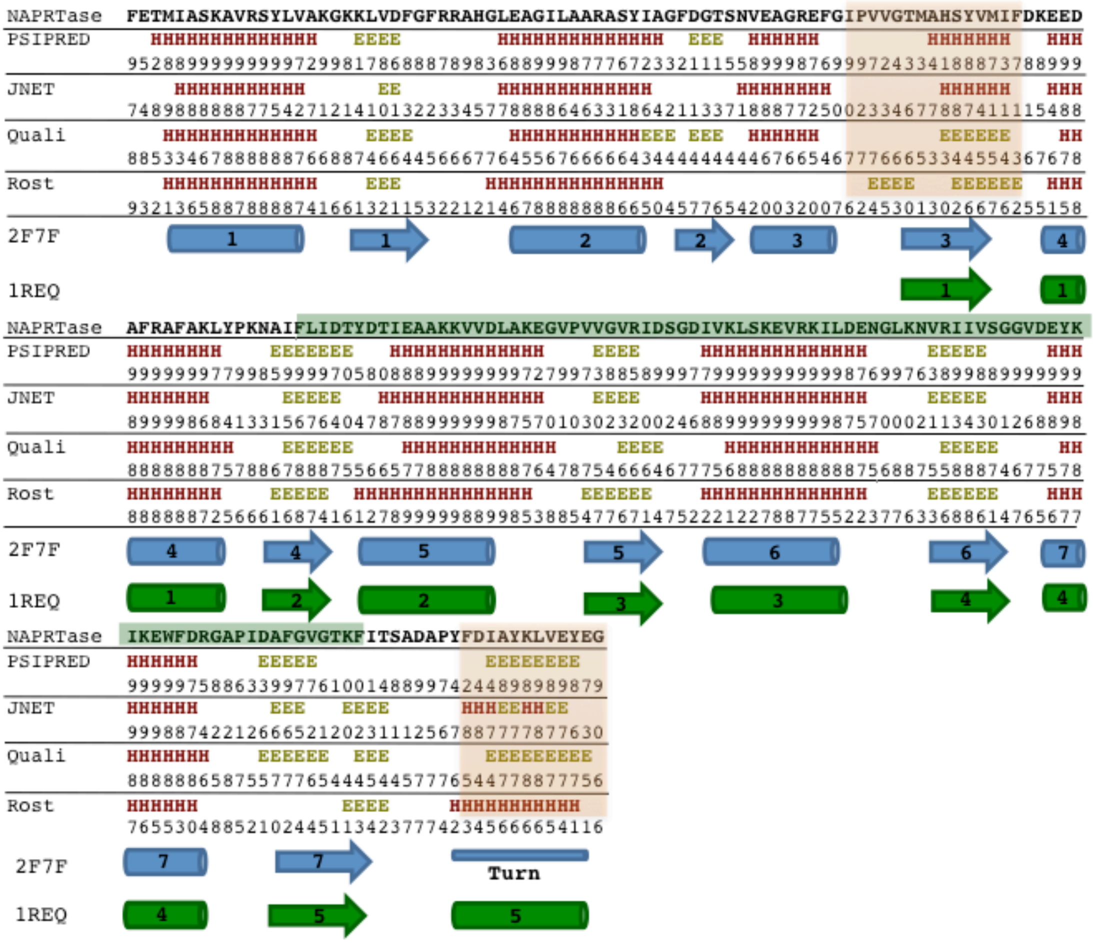


Fig. 17: Secondary structure prediction of the NAPRTase from *Aquifex aeolicus*. The prediction of 4 servers (PSIPRED, JNET, Quali, and Rost) gathered by the meta-server Quick2D (*68*), is depicted as alpha helices (**H**) and beta strands (**E**). The confidence values for the predictions are shown in numbers (range 0-9). The sequence of the NAPRTase was aligned to the canonical barrel domain from 2F7F (the secondary structure elements are shown in blue) and the B12-binding domain from 1REQ (the secondary structure elements are shown in green). We aligned the two folds using the NAPRTase sequence as anchor. Transparent orange squares denote plastic regions of the NAPRTase sequence, where the prediction of the secondary structure type is ambiguous. The NAPRTase is 42% identical with 2F7F over the whole alignment; while it is 29% identical with 1REQ over 85 residues (indicated with a shaded green bar).

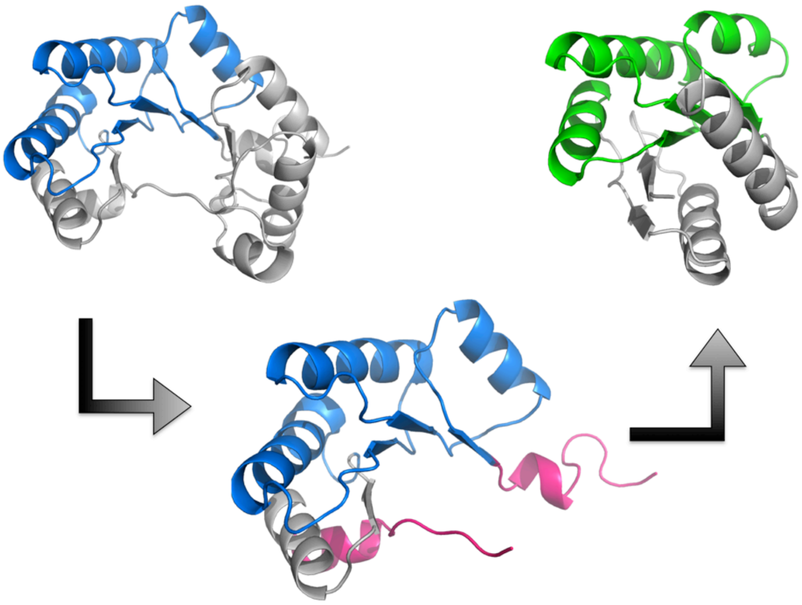


Fig. 18: Hypothetical evolutionary path of fold change between the ($\beta/\alpha)_7$-barrel 2F7F towards the flavodoxin-like fold B12-binding domain (1REQ). In blue and green are highlighted the $(\alpha\beta)_3$ elements shared by the two folds. In the lower model two plastic regions, which may facilitate the fold change path, are highlighted in pink.

The visual inspection of many CLANS maps provided us with a long list of interesting proteins; see section 8.4 from the appendix. Different aspects prompted us to select these targets; however, two PFAM families were the two most promising groups of proteins: a) the uncharacterized protein family UPF0004 appears in conjunction with a C-terminal domain of MiaB proteins, and b) the Radical SAM N-terminal family, which was already discussed previously. In section 8.4 from appendix we show many targets. It is indicated how similar the intermediate sequences are towards flavodoxin-like and $(\beta/\alpha)_8$-barrel fold structures. In addition, we show the gene identifier and the source organism. It is very difficult to evaluate whether any of the listed sequences will provide significant information about the evolution of the two superfolds under study. However, we can re-evaluate every hit before an experimental characterization.

Certain sequences were selected in function of many different factors: a) the similarity was high towards one fold but the length was more similar towards the other one, b) the protein was not annotated because there was no clear similarity towards any known function but it was similar to both folds, c) the length of the protein is compatible with a half barrel (100 residues) and shows high similarity towards both folds, d) we found many sequences from the same protein family, e) the topology suggested a duplicated flavodoxin-like fold in the barrel size, f) the protein is annotated with a function associated with a flavodoxin-like fold but it has the size of a $(\beta/\alpha)_8$-barrel. From the overall analysis of the candidates outlined here, we selected some proteins that were cloned in different versions. In the next chapter we discuss the experimental characterization of the candidates.

3.5 Experimental characterization of the intermediate candidates

The main goal of the experimental work was to obtain structural information in atomic detail from the intermediate candidates. Therefore, we selected several proteins to be experimentally characterized. For most of the hits, it was only possible to perform partial characterization steps. A summary of the data generated from 20 different proteins is provided in table 13. We determined the crystal structure of NTM0182, one of the intermediate proteins. The characterization procedure is fully outlined only for this variant.

Table 13: Summary of the data generated from the intermediate candidates experimentally tested. The amount of protein, in the soluble or insoluble fraction, is indicated by the number of (✓) symbols. Success or failure of the refolding protocol and crystallization attempts is indicated with (✓) or (X) respectively.

	Soluble	**Insoluble**	**Refolding**	**Monomer**	**Dimer**	**Multiple oligomerization states**	**Secondary structure content (CD spectra)**	**Enough yield to set crystal screening**	**Crystallized**
TM02	✓							✓	3.2 Å
CM01	✓	✓✓	✓	✓		✓	✓		
CM02	✓	✓✓	✓	✓					
SS01		✓	*X*						
SS02		✓							
SS03		✓	*X*						
PC01	✓	✓✓	*X*						
PC02	✓	✓✓	*X*						
PC03	✓	✓✓	*X*						
PC04	✓✓	✓✓	*X*		✓	✓	✓	✓	9.0 Å
PC05	✓	✓✓	*X*						
PC06	✓	✓✓	*X*						
HI01		✓	*X*						
CB01	✓	✓✓	✓						
CB02	✓	✓✓							
CT01		✓	*X*						
OB01		✓	*X*						
OB02		✓	*X*						
SM01	✓			✓			✓	✓	*X*

3.6 NTM1082 characterization

We cloned, expressed and purified the 128 N-terminal residues of the hypothetical protein TM0182 from *Thermotoga maritima* (NTM0182). In order to define the domain boundaries, we analyze both, a multiple sequence alignment and a consensus result obtained from various web servers (PRODOM, DOMPRED and PSIPRED). First, we obtained the gene from the hypothetical protein TM0182 by PCR from genomic DNA. We cloned the first 128 residues including a 6X-His-Tag.

The molecular weight of the protein is 14952.1 Daltons, and it has an isoelectric point of 6.29. We purified the protein by nickel affinity and size exclusion chromatography. The oligomeric state was characterized using analytical gel filtration and light scattering. In the purified protein solution, two concentration independent oligomerization states were identified corresponding to a monomer and a dimer (see figure 19 for details). The stability of both states can be explained by analyzing the atomic model determined. We will discuss this finding later on this dissertation.

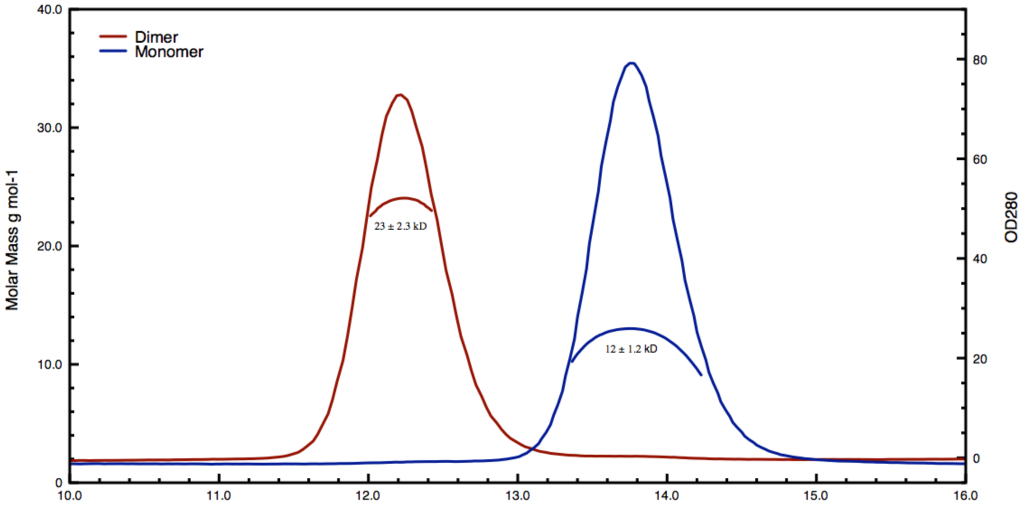


Fig. 19: Light scattering and gel filtration curves of monomeric (blue) and dimeric (red) NTM0182. The experimentally determined molecular weight is indicated.

Analytical gel filtration was performed using a column Superdex 75 10/300 GL column (Amersham Pharmacia) and was coupled to a light scattering device for subsequent analysis. Multiangle static laser light-scattering experiments (MALLS) were done online with analytical size-exclusion chromatography using miniDAWN TREOS and Optilab rEX instruments (Wyatt Technologies) and the associated software (AstraV) for molecular weight determination. The protein (0.3 mg/ml) was eluted at a flow rate of 0.8 ml/min in 50 mM Tris, 300 mM KCl (pH 7.5) and the apparent size of the two peaks was analyzed by dynamic light scattering. The molar mass (g/mol) was determined: monomer 1.292e+4 (error of 0.8%), dimer 2.382e+4 (error of 1%).

The results from far-UV circular dichroism suggested that, in agreement with previous bioinformatic predictions, NTM0182 displayed a mixed content of beta/alpha secondary structural elements. Both oligomeric states show a very similar curve (Figure 20).

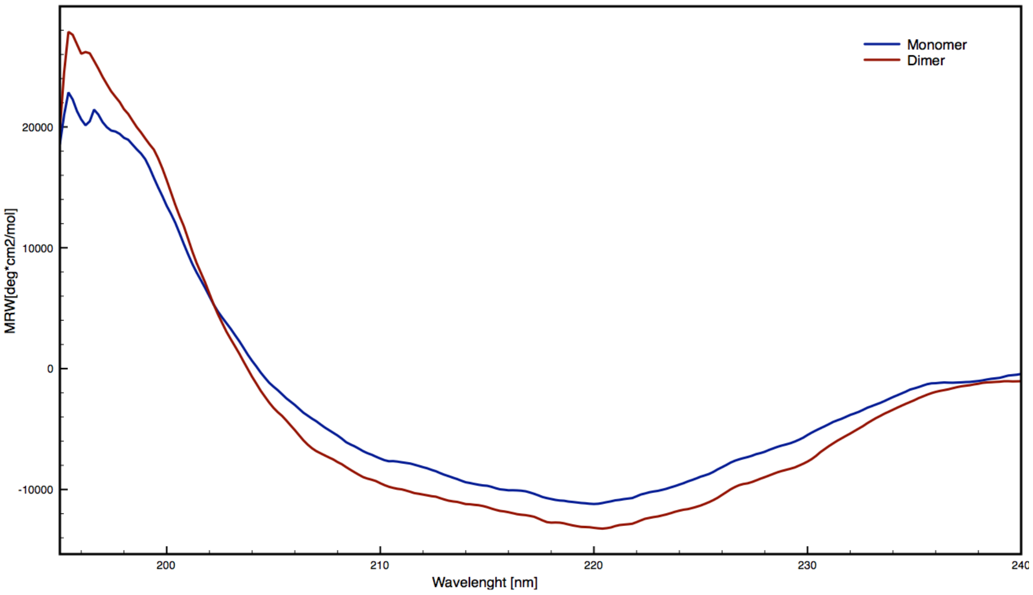


Fig. 20: CD spectra of monomeric (blue) and dimeric (red) NTM0182 CD spectra were recorded with a JASCO model J-810 spectropolarimeter. The measurements were performed using protein concentration of 0.1 mg/ml in 20 mM Tris and 100 mM KCl (pH 7.5) at room temperature.

The tertiary structure was evaluated by fluorescence spectroscopy (figure 21). The dimeric protein shows a maximum at 320.5 nm while the monomer curve displays a maximum at 340 nm. Also, the intensity of both curves is different, being lower for the monomer, although the concentration of both samples is the same. Overall, the data suggests a better shielding of aromatic residues in the dimeric arrangement.

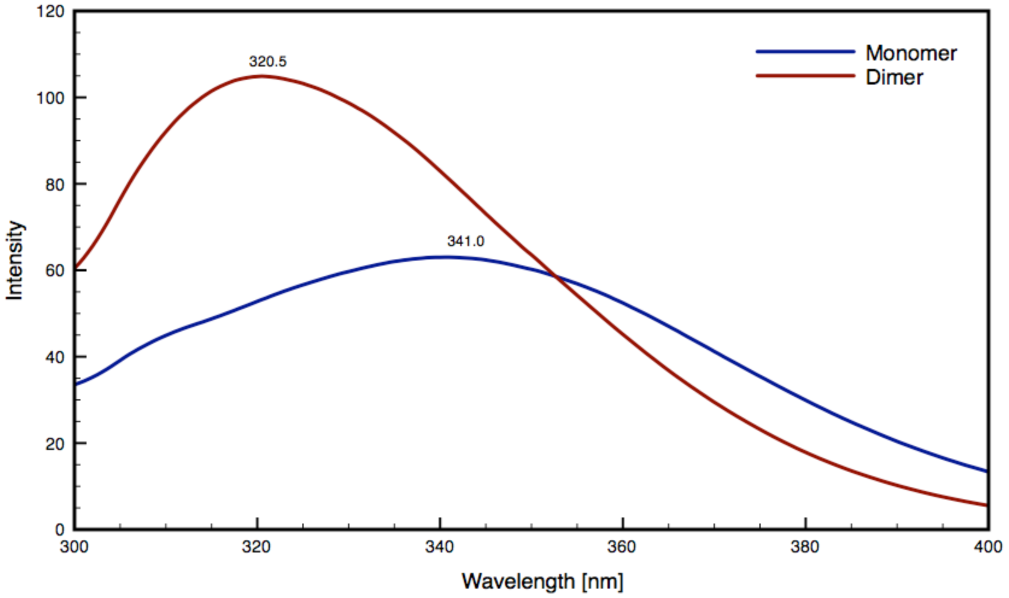


Fig. 21: Fluorescence spectroscopy curves of monomeric (blue) and dimeric (red) NTM0182. The measurements were carried out with a JASCO FP-6500 spectrofluorometer; the experiment was performed using protein concentration of 0.1 mg/ml in 20 mM Tris and 100 mM KCl (pH 7.5) at room temperature. The monomer spectrum has a maximum at 341 nm, while the dimer spectrum has a maximum at 320 nm.

The biophysical characterization of NTM0182 shows a well-folded protein domain, with a mixed alpha/beta content, that adopts two concentration-independent oligomerization states. Nonetheless, the data provided here does not aid to understand the evolution of the two folds under study. Therefore, we set crystallization trials to produce an atomic model of NTM0182.

The x-ray crystal structure of NTM0182 was determined by experimental phasing. We selected this methodology instead of molecular replacement because the average sequence identity shared between NTM0182 and any known protein structure was lower than 18%; moreover, we wanted to avoid bias towards any fold while solving the phase problem.

A crystallization screening produced hits in different conditions; the crystals were needle-like shaped in different sizes. We determined that NTM0182 crystallizes in an orthorhombic lattice of the space group C2221, the solvent content of the crystals was 58%, and there were six monomers per asymmetric unit. The crystals diffracted up to 3.2 Å resolution. A double SAD (single wavelength-anomalous diffraction) experiment was performed where two anomalous datasets, coming from different crystals soaked in platinum, were collected. We merged these datasets to achieve enough phasing power to determine an initial experimental map. Several cycles of density modification, model building, and refinement were carried out to obtain a final refined model of NTM0182 (R-work of 0.24, R-free of 0.28). See table 14 for full X-ray, phasing and refinement statistics for NTM0182

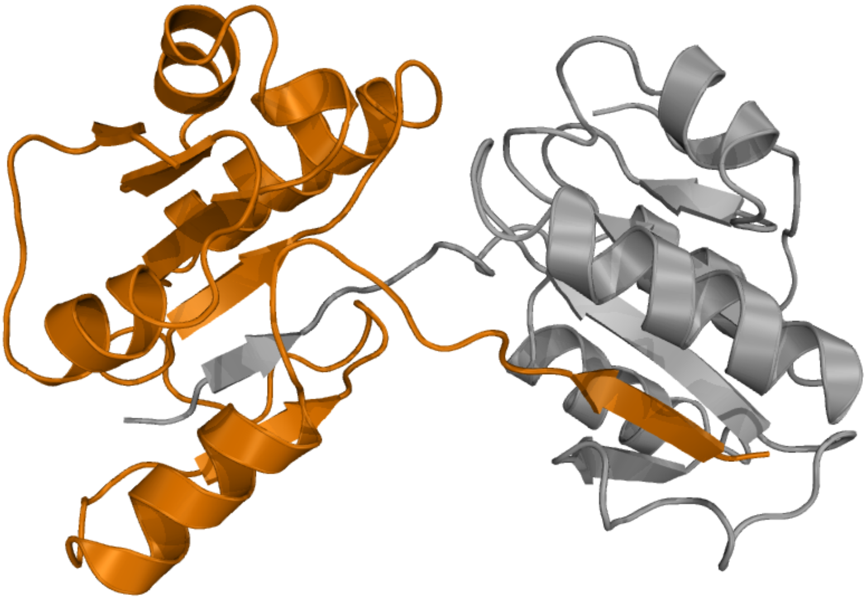


Fig. 22: The X-ray crystal structure of NTM0182 was determined by experimental phasing. The monomers in the asymmetric unit swapped the first N-terminal β-strand forming a close dimeric arrangement. The chains D (orange) and F (gray) are shown in cartoon representation.

Table 14: X-ray, phasing and refinement statistics for NTM0182

Data collection	
Wavelength (Å)	1.069800
Space group	C222(1)
Cell dimensions (Å, °)	a=102.05,b=145.30, c=141.71 a=90.0, b=90.0, γ=90.0
Resolution (Å)	39.6-3.19 (3.27-3.19)
Unique reflections	33,708 (2,241)
Redundancy	6.3(5.7)
Completeness %	99.1%(88.6%)
Rmerge %	14.7 (83.1)
I/sigma (I)	15.01(2.02)
Wilson B factor	67.8
Phasing statistics	
No of Platinum sites	2
Phasing power Anomalous	0.982 (0.205)
$\boldsymbol{R}_{\mathrm{cullis}}$	0.966
Refinement statistics	
Space group	C222(1)
Resolution (Å)	39.6-3.19 (3.3-3.19)
Figure of merit	0.77 (0.70)
Phase error	29.82 (36.55)
R_{cryst}, %	0.2465 (0.33)
R_{free}, %	0.2881 (0.37)
Mean B-value (Å^2)	78.5
Nonhydrogen atoms	5246
Number of water molecules	29
Number of platinum molecules	2
rmsd of bond length (Å^2)	0.004
rmsd of angle (°)	1.005
Model quality	
Clashscore, all atoms	8.09 (95th percentile)
MolProbity score (*69*)	2.02 (100th percentile)
Poor rotamers	1.0%
Ramachandran outliers	0.31%
Ramachandran favored	90.22%
Cβ deviations	3
Residues with bad bonds	0.00%
Residues with bad angles	0.06%

Three layers of secondary structural elements compose the NTM0182 structure: a six-stranded β-sheet (order 213456) sandwiched by two α helices on one side, and three α helices on the other side (figure 22). The arrangement of the monomers in the asymmetric unit was strikingly unusual; the first ten residues (β1 + 3 residues) from each monomer were swapped. This beta-strand swap produced a closed dimeric arrangement. A detailed observation of the components of the asymmetric unit showed that the chains did not superpose well. Instead, the angle formed by the residues after the swapped region from each monomer deviated by several degrees. The closed dimeric arrangement in the crystal structure correlates very well with the concentration independent oligomerization state of the protein in solution.

Three chains displayed C-β deviations (TRP12 from chain B, ASN10 and TRP12 from chain D) that could not be fixed, despite many cycles of rebuilding and refinement. These positions are localized where the swapped β-strands point away from the monomers. The swapping of the β-strands may be the cause of this unusual atomic arrangement.

To further perform structural comparisons it was necessary to define an operational monomer: we took β-strand 1 from chain F and the rest of the secondary structural elements from chain D. The topology of NTM0182 did not precisely match any fold in the most widely used classification systems (SCOP and CATH); therefore, it could be defined as a novel fold.

3.7 Sequence and structural based comparison of NTM0182 with the flavodoxin-like and the $(\alpha\beta)_8$-barrel folds.

We aligned (guided by profile-profile pairwise sequence alignments) the NTM0182 structure with the Inosine monophosphate dehydrogenase from *Streptococcus pyogenes* [PDB 1ZJF, $(\alpha\beta)_8$-barrel fold) and the glutamate mutase (B12-binding domain) from *Clostridium cochlearium* (PDB 1I9C, Flavodoxin-like fold). The protein sequence from NTM0182 was initially detected using the sequence of 1I9C as query for a HHsenser search.

We followed sequence based information, in order to guide the structural superpositions, because many members of the SCOP α/β class show structural similarities that may be the result of convergence and not necessarily due to common ancestry (data not shown). A detailed inspection of these superpositions revealed a conserved (in sequence and structure) αβαβ element among the three different folds (figure 23).

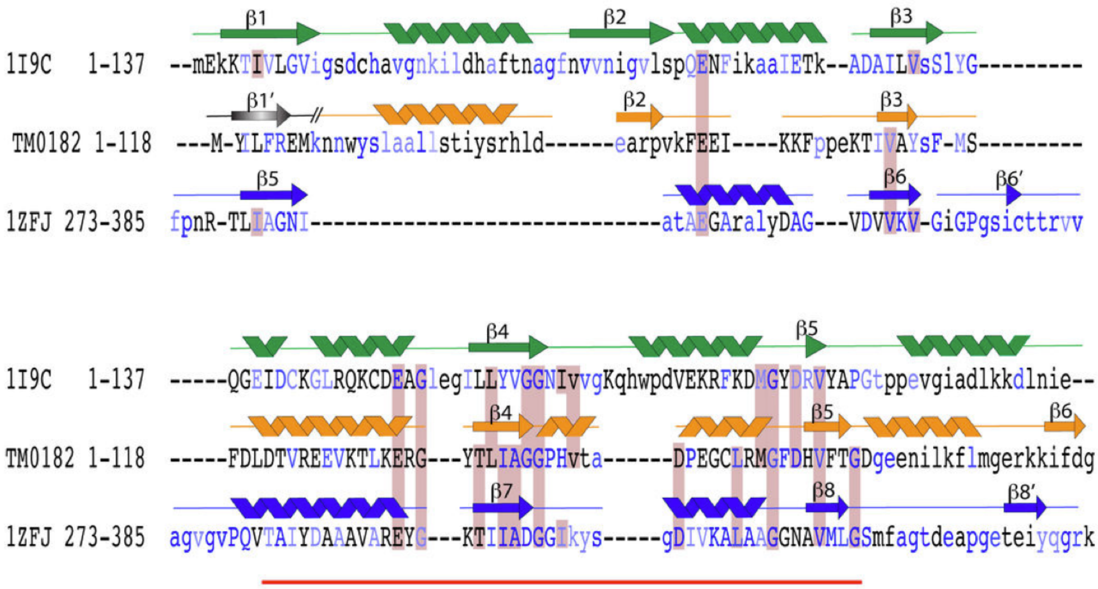
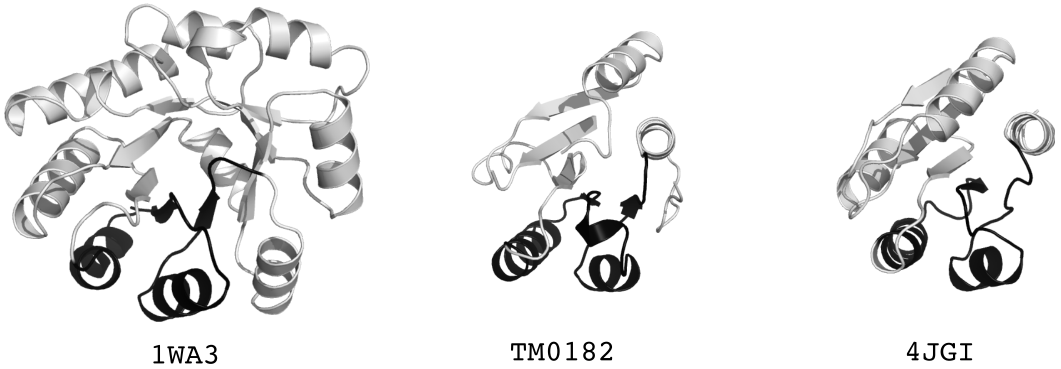

β1
β2
β3
1I9C 1-137 --mEkKTIVLGVigsdchavgnkildhaftnagfnvvnigvlspQENFikaaIETk--ADAILVsSlYG---------
β1′
β2
β3
TM0182 1-118 ---M-YILFREMknnwyslaallstiysrhld------earpvkFEEI----KKFppeKTIVAYsF-MS---------
β5
β6
β6′
1ZFJ 273-385 fpnR-TLIAGNI-----------------------------atAEGAralyDAG---VDVVKV-GiGPgsicttrvv
β4
β5
1I9C 1-137 -----QGEIDCKGLRQKCDEAGlegILLYVGGNIvvgKqhwpdVEKRFKDMGYDRVYAPGtppevgiadlkkdlnie--
β4
β5
β6
TM0182 1-118 -----FDLDTVREEVKTLKERG---YTLIAGGPHvta------DPEGCLRMGFDHVFTGDgeenilkflmgerkkifdg
β7
β8
β8′
1ZFJ 273-385 agvgvPQVTAIYDAAAVAREYG---KTIIADGGIkys-----gDIVKALAAGGNAVMLGSmfagtdeapgeteiyqgrk

Fig. 23: A structure-based sequence alignment of the of NTM0182 (orange) with the flavodoxin-like protein glutamate mutase from *C.cochlearium* (PDB 1I9C, in green) and the $(\beta\alpha)_8$-barrel inosine monophosphate (IMP) dehydrogenase from *S.pyogenes* (PDB 1ZFJ, in blue). The alignment reveals an area of highest similarity around α3β4α4β5 indicated by the red bar. The amino acid sequences are shown with the secondary structural elements indicated above. Increasing sequence conservation within each family is indicated through more intense blue shading of the 1-letter code. Capital letters denote structurally aligned residues. Identical residues between the three proteins are highlighted by reddish background.

Interestingly, the structure-based sequence alignment shows that the highest similarity of NTM0182 to either the flavodoxin-like or the (βα)8-barrel fold is confined in a 42 residues long fragment that encodes the α3β4α4β5 element (figure 23).

While the NTM0182 domain aligns upon superposition with only 13 and 19% over 67 superimposed residues to glutamate mutase and IMP dehydrogenase, respectively, the α3β4α4β5 element from NTM0182 shows a higher sequence identity of 23% and 28% over 40 aligned residues with the corresponding fragments of glutamate mutase and IMP dehydrogenase. On the contrary, the proteins glutamate mutase and IMP dehydrogenase share only 15% sequence identity in both, the local and the global alignment. The structural alignment may further suggest that the homology among the three different folds is restricted to a smaller fragment, e.g. the conserved $(\alpha\beta)_2$ element.



```
1WA3      11-46     KIVAVLR---ANsveeak------------------------ekalaVF----EGGv---HLIEITFTvP-
                      ||                                                         ||  |*
TM0182    1-56      MYILFRE---MK-NNWY--SLAALLSTiysrhldVEARPV----KFEEI----KKFPpeKTIVAYSFMSF-
                      *|                   |  |        *|  |        *    *       **  * |
4JGI      90-152    AKIVLATvegDLhDIGKniFRTMAEASg------FEVFDLgidvPVKIIvdkvKEVN--PEIVGLSGVLTl

1WA3      47-110    DADTVIKELSFLKEK-GA---IIGAGT--VTsveqCr-kAVE-SGAEFIVSpHldeeisqfckekgVFYMPG
                    *|***  *|  ***| *     | **   **         |   * | | |   |* *  *          *
TM0182    57-120    DLDTVREEVKTLKER-GY---TLIAGGphVTa---DpegCLR-MGFDHVFTgDGeENILKFLMGErKKIFDG
                     **||** *  **         |*|**  *             |* *   *      |*
4JGI      153-211   ALDSMRETVDALKAEgLRndlKVIIGGvpVNe---N---VCQrVGADDFST-NA-ADGVKICQRW-vg----
```

Fig. 24: Structure-based sequence alignment of the N-terminal domain of TM0182 with the flavodoxin-like B12-binding from *Desulfitobacterium hafniense* (PDB-id: 4JGI) and the $(\beta\alpha)_8$-barrel class I KDPG aldolase from *Thermotoga maritima* (PDB-id: 1WA3). The $(\alpha\beta)_2$ fragment with the higher identity/similarity is colored in black in the alignment as well as in the structural models (depicted as cartoon), while the rest is in gray. Capital letters denote structurally aligned residues. The scores for the superpositions are shown in table 14.

The high sequence identity of the element $(\beta\alpha)_2$ among all the folds, could be seen as suggestive evidence of common ancestry. However, the identities would lie below certain strict empirical thresholds to support homology between proteins based on fragment length and sequence identity (*70*). We therefore searched for proteins with even higher sequence identities and identified a recently released flavodoxin-like structure (PDB-id: 4JGI, SCOP: c.23.6) as well as a $(\beta\alpha)_8$-barrel structure (PDB-id: 1WA3, SCOP: c.1.10) that both share 37% sequence identity within the $(\beta\alpha)_2$ fragment of TM0182 (alignment in figure 24).

The data generated by the HMM-HMM comparisons combined with the previously discussed structural superpositions, significantly diminish the possibility of independent origin of the (βα)8-barrel and the flavodoxin-like fold.

Table 15: Scores for alignment in figure 24.

Structurally aligned	**SeqID Global**	**RMSD Global**	**SeqID Local**	**RMSD Local**
1WA3 *vs.* TM0182	24 %	2.5 Å	37 %	2.3 Å
4JGI *vs.* TM0182	20 %	2.3 Å	37 %	2.0 Å
1WA3 *vs.* 4JGI	9 %	2.7 Å	18 %	2.8 Å

3.8 The $(\alpha\beta)_2$ element can be structurally superposed onto different $(\beta/\alpha)_8$-barrel and flavodoxin-like fold superfamilies.

Previously, different research groups (*27, 71*) proposed an evolutionary scenario in which HisA and HisF, histidine biosynthesis enzymes, evolved through duplication and fusion of a gene encoding a half-barrel ancestor. The Nagano group (*31*), in a very extensive work on the $(\beta/\alpha)_8$-fold, found a common G-X-D motif in the even loops of the barrels (α1-β2; α3-β4; α5-β6; α7-β8). This structural feature suggested a 4-fold duplication of an ancestral $(\alpha\beta)_2$ motif. In our global analysis, using state of the art tools for homology detection, we confirm the same relationships among $(\beta/\alpha)_8$-barrel fold superfamilies (*30, 31*) that were previously established; moreover, we add new sequence-based links to additional superfamilies.

The 4-fold or 2-fold duplication has been then repeatedly hypothesized as the source of the approximate 8-fold structural symmetry in the $(\beta/\alpha)_8$-barrel fold. Söding and co-workers, with the development of the HHrep software (*72*), provided sequence-based evidences that support the evolution of the $(\alpha/\beta)_8$ barrel fold by duplication and fusion of smaller (αβ) motifs. They found both, a strong (P-value 9.7x10-13) two-fold symmetry, and a weaker 4-fold symmetry (P-value 5.1x10-04) in HisF (PDB 1THF). Other $(\beta/\alpha)_8$-barrel fold superfamilies also displayed 4-fold symmetry: the KDPG aldolase (PDB 1fg0; SCOP c.1.10.1) displayed a weak four-fold symmetry (P-value 5.4x10-7), and the phosphoenolpyruvate mutase (PDB 1s2w , SCOP c.1.12.7). Lastly, a three-fold symmetry was detected in the inosine monophosphate dehydrogenase (PDB 1ZFJ, SCOP c.1.5.1). Together, these results suggests that the $(\beta/\alpha)_8$ barrel fold may have arisen by a 4-fold duplication of a $(\beta/\alpha)_2$ module.

Our rationale indicates that a homologous $(\beta/\alpha)_2$ element, which links the flavodoxin-like with the $(\beta/\alpha)_8$-barrel fold, should also connect the $(\beta/\alpha)_8$ barrel fold superfamilies that were either established as having common ancestry, or proven to display symmetric properties. In table 16 we summarize the results of multiple structural superpositions, using the α3β4α4β5 element**,** of NTM0182 with eleven $(\beta/\alpha)_8$-barrel fold superfamilies and one flavodoxin-like fold superfamily. Our results are completely consistent with previous findings.

Table 16: Superposition of the NTM0182 αβαβ element on different (βα)$_8$-barrel fold structures.

SCOP Superfamilies	**SCOP (PDB)**	**RMSD**	**% Seq. ID**	**Aligned length (αβαβ)**
Triosephosphate isomerase (TIM)	c.1.1.1(1W0M)	2.0	22.0	41
Ribulose-phoshate binding barrel	c.1.2.1 (1THF) c.1.2.5 (3IGS) c.1.2.1 (1KA9)	2.1 1.8 2.0	27.3 29.7 29.3	40 37 41
Thiamin phosphate synthase	c.1.3.1 (1XI3)	2.2	30.0	30
FMN-linked oxidoreductases	c.1.4.1(1GOX)	2.2	25.0	44
Inosine monophosphate dehydrogenase (IMPDH)	c.1.5.1 (1ZFJ)	2.0	27.5	40
Aldolase	c.1.10.1 (1WA3)	2.7	31.7	41
Phosphoenolpyruvate domain	c.1.12.1 (1E0U)	2.8	27	37
Nicotinate/Quinolinate PRTase C-terminal domain-like	c.1.17.1 (1QPO)	1.6	18.0	39
PLC-like phosphodiesterases	c.1.18.1 (2PZ0)	1.4	31.4	35
(2r)-phospho-3-sulfolactate synthase ComA	c.1.27.1(1QWG)	2.3	28.6	42
GlpP-like	c.1.29.1.1 (1VKF)	1.6	23.5	34
Cobalamin (vitamin B12)-binding domain	c.23.6.1(2XIJ)	2.0	28.9	38

The fragment α3β4α4β5 has an equivalent structural match with high sequence identity and low RMSD in 11 different (βα)$_8$-barrel fold superfamilies. These superfamilies were already proven to be homolgous (e.g Ribulose-phoshate binding barrel and Thiamin phosphate synthase) or have displayed a three-fold repeat pattern (Inosine monophosphate dehydrogenase).

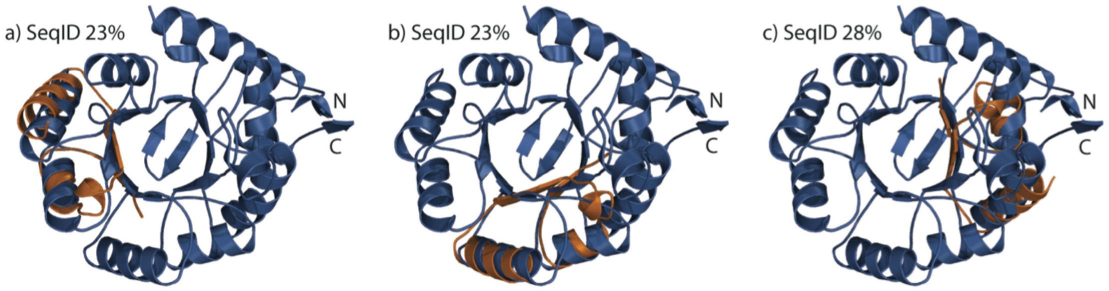
3.9 The $(\alpha\beta)_2$ element can be shifted and superposed three times onto the Inosine monophosphate dehydrogenase structure.

The Inosine monophosphate dehydrogenase previously showed a three-fold repeat pattern (PDB 1ZFJ; SCOP c.1.5.1). We tested whether the detected conserved fragment would match the three-fold repeat in 1ZFJ. For this, we superposed and shifted this element three consecutive times onto the barrel. The results of the superpositions are summarized in table 16.

Table 17: Three consecutive structural superpositions of the NTM0182 αβαβ element on the 1ZFJ structure

Superposition range on 1ZFJ	**RMSD (Å)**	**% Seq. id (similarity)**	**Aligned length (αβαβ element)**
GLN83 — ASP 253 Conserved motif 246-248 GXD identical	2.1	23.32 (43)	37
SER259 — GLY303 Conserved motif 296-298 GXD identical	1.8	23.31 (42)	38
VAL323 — GLY 366 Conserved motif 296-298 GXN not conserved	2.0	27.50 (32)	40

Remarkably, the $(\beta\alpha)_2$ element can be superposed with significant high scores (when considering comparisons of proteins from different folds) over three consecutive regions (corresponding to a quarter barrel) onto 1ZFJ. These superpositions also match two times (only the glycine is conserved the third time) the conserved GXD motif that was interpreted as evidence for the $(\beta\alpha)_2$ modular unit from which the $(\beta\alpha)_8$ barrel fold might have originated.



```
a)    81-TEQAEEVRKVKRS//LLVAAAVGvtsdTFeRAEALFEAGADAIVID-253
      59-DTVREEVKTLKERg-YTLIAGGPh--vTA-DPEGCLRMGFDHVFTG-100

b)   260-AGVLRKIAEIRAHFpnRTLIAGN--IATaeGARALYDAGVDVVKVG-303
      59-DTVREEVKTLKERG--YTLIAGGphVTA--DPEGCLRMGFDHVFTG-100

c)   323-VTAIYDAAAVAREyGKTIIADGG--IKysgDIVKALAAGGNAVMLG-366
      59-DTVREEVKTLKER-GYTLIAGGPhvTA---DPEGCLRMGFDHVFTG-100
```

Fig. 25: The conserved α3β4α4β5 element can be superposed three consecutive times on 1ZFJ (Inosine monophosphate dehydrogenase).

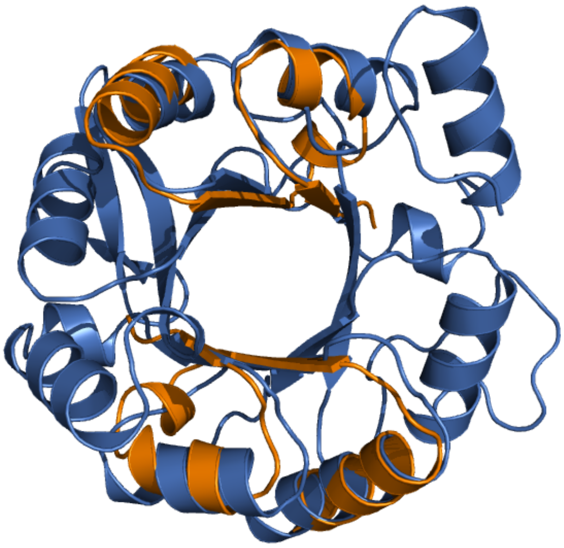


Fig. 26: The conserved α3β4α4β5 element can be superposed two times on 1KA9 (HisF). The scores are summarized in table 18.

Table 18: Structural scores and sequence identity for superpositions in figure 26

Superposition range on 1ka	**RMSD (A)**	**% Seq. id (similarity)**	**Aligned length ($\alpha\beta\alpha\beta$ element)**
ALA54 — ASN 103 Conserved motif 96-98 GAD identical	2.0	29.3 (40)	41
ASP183 — SER225 Conserved motif 218-220 GAE not identical but similar	1.7	21.6 (35)	37

HisF (e.g PDB 1ka9) has been the most mentioned example of a 2-fold symmetric $(\beta\alpha)_8$-barrel fold. Therefore, we expected to be able to superpose with good scores the conserved quarter barrel at least twice. Indeed, the $(\beta\alpha)_2$ element was superposed two times on the structure of HisF (1KA9), the imidazole glycerol phosphate synthase subunit, with low RMSD and high sequence identity. Moreover, we found good agreement on the superposition of the conserved GXD (in this case the motif was GAD) motif present in most $(\beta\alpha)_8$-barrel fold structures: the motif was completely aligned on the first half, and the second motif showed a glutamic acid instead of the canonical aspartic acid (figure 26 and table 18). Thus, our data supports the model of evolution previously proposed for the $(\beta\alpha)_8$-barrel fold, where a 4-fold duplication and fusion of a quarter barrel yielded the contemporary $(\beta\alpha)_8$-barrel fold.

3.10 Functional constraints may be the basis of the high conservation of the $(\alpha\beta)_2$ element.

The GG motif for instance (involved in the binding of the nucleotide-region of the B12 molecule) is located in equivalent positions in a structural alignment between 1BMT and NTM0182 (11 identical residues; 86 C-alpha carbons and 2.2 Å RMSD). The structure of NTM0182 showed a dimer configuration where the first beta-strand (residues 1-8) from each chain is swapped between two interacting monomers. W11 is the first residue that packs back against the source monomer that swapped the first N-terminal beta strand (see figure 27).

In the surrounding regions of W11 in all monomers present in the asymmetric unit, the geometrical scores for NTM0182 are not optimal (rotamer occurrence, length and angles of the bonds). The structural superposition in figure 27 places the indole ring of W11 almost completely aligned in the same plane with the indole ring of the nucleotide from the B12 in 1BMT.

The NTM1082 β-strand swapping may be a crystallographic artifact caused by the missing C-terminal domain in the following way: the lateral chain of W11 may be mimicking the binding of a nucleotide-containing substrate on NTM0182 and this fact triggers the unusual packing in this area.

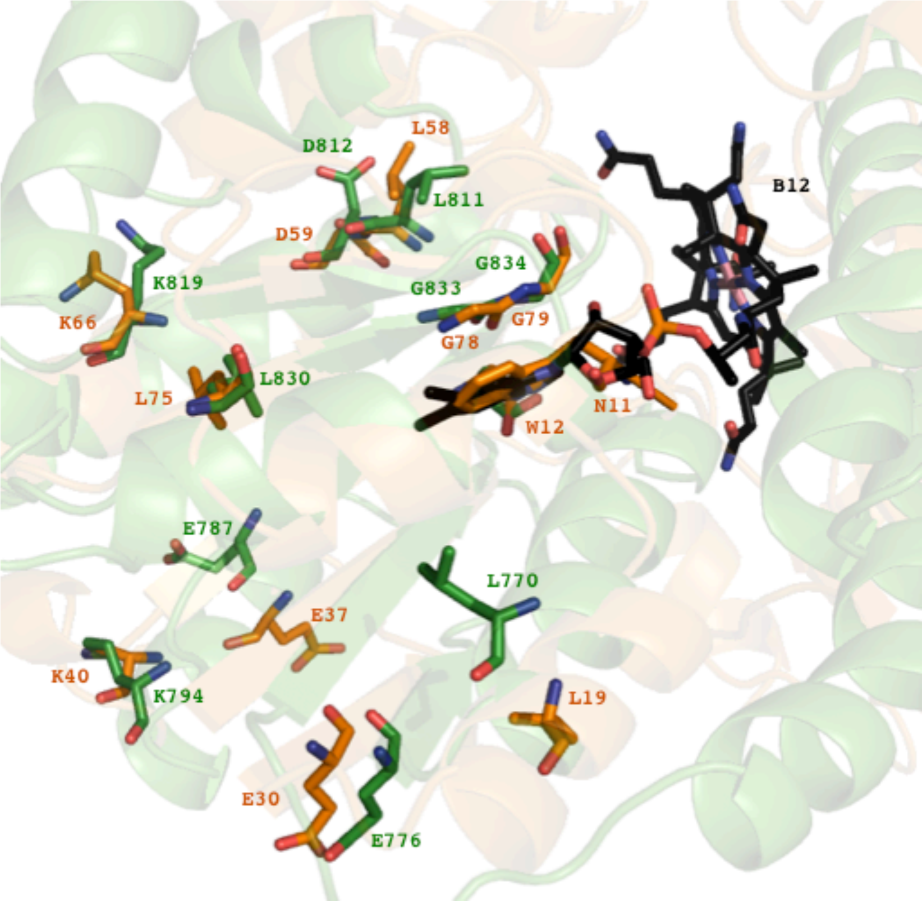


Fig. 27: Structural alignment of 1BMT (green) and NTM1082 (orange). 11 identical residues between both structures identified in the alignment are highlighted in stick representation. B12 (1BMT ligand) is shown in black sticks. The lateral chain of W12 from NTM0182 (stick orange representation) aligns with the indole ring of B12.

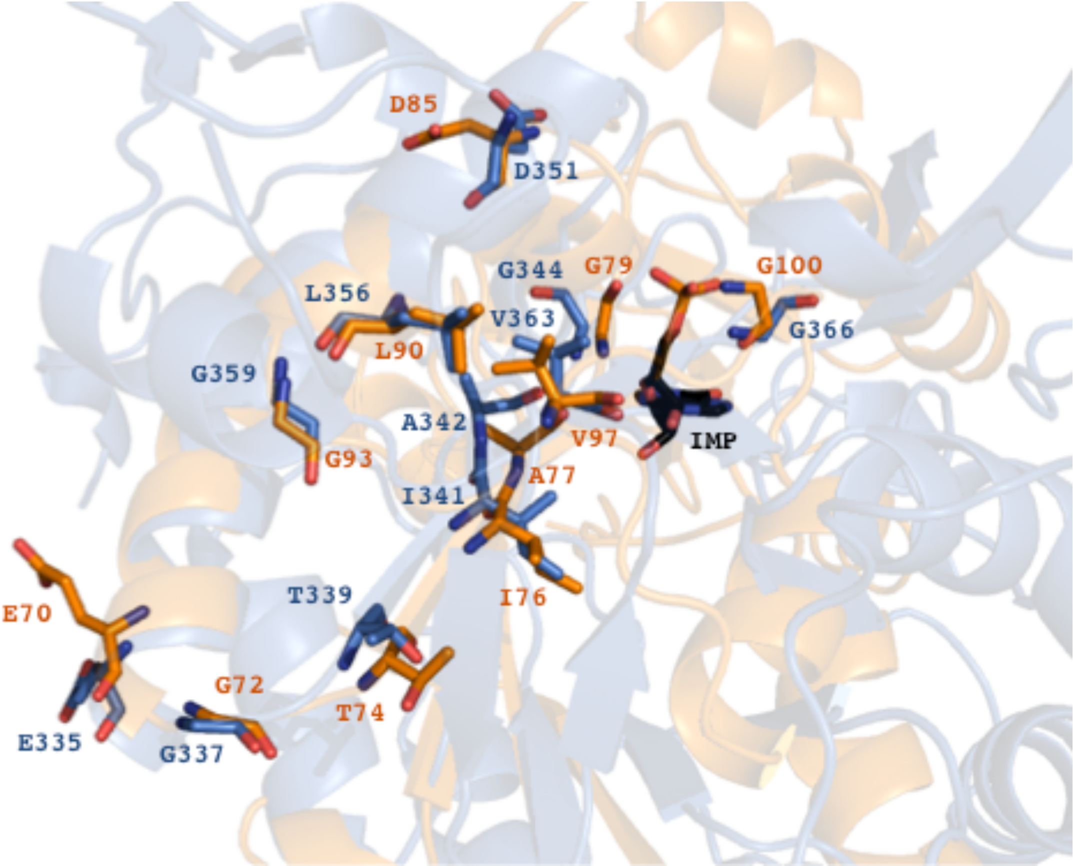


Fig. 28: Structural alignment of 1ZFJ (blue) and NTM1082 (orange). 11 identical residues between both structures are highlighted in stick representation. IMP (Inosinic acid, the ligand of 1ZFJ) is depicted in black stick model. Two glycines (phosphate-binding site in many (β/α)$_8$-barrel structures) are aligned in both structures and lie in range (> 3.5A) of establishing interactions with the phosphate group of the ligand.

In contrast to the previously explored superposition between the flavodoxin-like fold and NTM0182, the identities (11 residues) found between 1ZFJ and NTM1082 are confined to the αβαβ element (40 C-alpha carbons and 1.7 A RMSD). The αβαβ module of 1ZFJ and NTM1082 are more identical between each other than with 1BMT, which in turn displays a more similar overall fold to NTM0182. This fact emphasizes the intermediate nature of NTM0182. Two glycines that make contacts with the ligand (inosinic acid) in 1ZFJ are conserved in the structural superposition (figure 28). Both are very close to the nucleotide part of the inosinic acid, particularly near to the phosphate ion.

In summary, we may conclude that the conservation of the αβαβ element among the three protein folds is preserved due to functional reasons. The maximal sequence identity, and structural similarity, is confined to the region where the nucleotide binds in both, the flavodoxin-like fold (B12) and the $(\alpha\beta)_8$-barrel fold (inosinic acid). In consequence, we could hypothesize that the three folds are binding ligands with phosphate moieties. The conserved αβαβ element will be involved in the interaction with the common phosphate moieties, and the rest of the fold evolved to accommodate divergent areas of the ligands.

Having found evidences of common ancestry between the $(\alpha\beta)_8$-barrel and the flavodoxin-like fold, we proceeded to explore the rest of the α/β class in SCOP. Our approach combined profile-profile comparisons with structural superimpositions.

3.11 Sequence-based exploration of the α/β class in SCOP.

The profile-profile comparisons of the $(\alpha\beta)_8$-barrel and flavodoxin-like fold revealed striking local structural similarities. We then envisioned that by comparing Profile Hidden Markov Models (HMM) representing all the α/β class in SCOP we could generate a database of interchangeable sub-domain fragments from different folds. Potentially, this database could be use as a framework to engineer novel functionalities that are present in different folds and could be combined in a single scaffold. As starting point, we aim to produce a well-folded protein starting by using fragment from divergent folds.

We took all the profiles that represent the structures classified in SCOP70 as α/β folds (147 folds) and we compared all against all. By this we generated a database of HHsearch outputs that contained the result of the comparisons. The database contained a huge amount of data; therefore we decided to only explore the most ancient connections.

To determine which were the most ancient folds we refer to an extensive work from Caetano-Anollés and coworkers (*16*). This work is based on a genomic demography involving hundreds of genomes. They measured the frequencies of occurring of protein folds in individual genomes as a phylogenetic character that describe how popular folds are in nature. They derived a phylogenetic tree in which the 3-helical bundle is the most ancestral fold followed by the $(\alpha\beta)_8$-barrel and the Rossmann fold. We therefore started looking for high-scoring pairwise connections established among the folds listed in table 19.

Table 19: The most ancestral folds in nature, the table was adapted from (*16*)

SCOP label	Fold
c.37	P-loop containing nucleoside triphosphate hydrolases 3 layers with α/β/α arrangement, parallel or mixed β-sheets of variable sizes
a.4	DNA/RNA-binding 3-helical bundle Core: 3-helices; closed or partly opened bundle, right-handed twist; up-and-down
c.1	(β/α)$_8$-barrel Closed barrel with parallel β-sheet and strand order 12345678
c.2	NAD(P)-binding Rossmann-fold domains Core: 3 layers in α/β/α arrangement; parallel β-sheet of 6 strands, order 321456
d.58	Ferredoxin-like Core: 3 helices; closed or partly opened bundle, right-handed twist; up-and-down
c.23	Flavodoxin-like 3 layers with α/ β /α arrangement; parallel β-sheet of 5 strands, order 21345
c.55	Ribonuclease H-like motif 3 layers with α/β/α arrangement; mixed β-sheet of 5 strands, order 32145 with strand 2 antiparallel to the rest
b.40	OB-fold Closed or partly opened barrel, with greek-key motif
c.66	S-adenosyl-L-methionine-dependent methyltransferases Core: 3 layers with α/β/α arrangement; mixed β-sheet of 7 strands, order 3214576 with strand 7 antiparallel to the rest

The Virulence factor MviM from *Escherichia coli* (SCOP d1tlta1 c.2.1.3, Rossman fold) was aligned with the KDO8P synthase from *Haemophilus influenza* (SCOP d1o60a_ c.1.10.4 (αβ)$_8$-barrel fold). 74 columns were aligned with a probability of 76.29 and E-value of 0.43. Having found the homologous region with HHsearch we then structurally superimposed 62 C-alpha carbons with an RMSD of 2.31 and 4.680 Z-score (figure 29). The Z-score indicates how statistically significant is the superposition

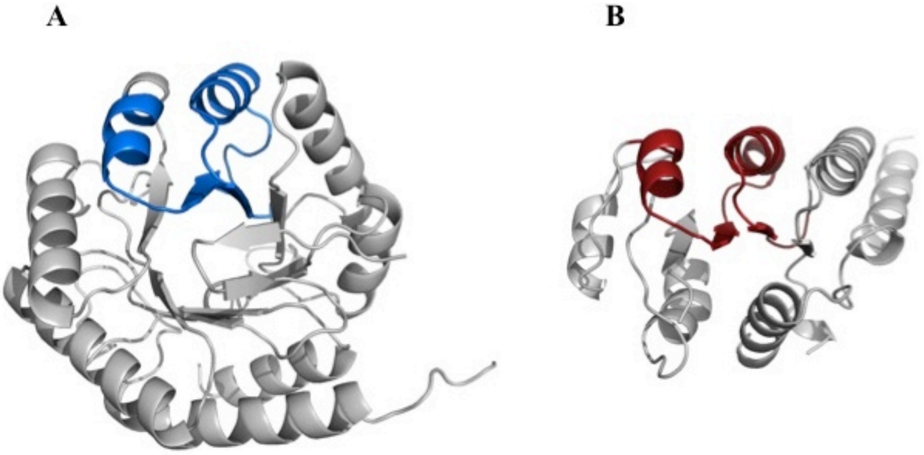


Fig. 29: A sequence-based profile-profile alignment between different folds indicates structurally conserved fragments. A) The KDO8P synthase and B) Virulence factor MviM. The aligned region is highlighted in color.

The NAD(P)-binding Rossmann-fold (SCOP c.2) also hit with high HHsearch probability another ancestral fold, the S-adenosyl-L-methionine-dependent methyltransferases (SCOP c.66). The bacterial secondary alcohol dehydrogenase from *Clostridium beijerinckii* (SCOP c.2.1.1) was aligned to the hypothetical protein Ta0852 from *Thermoplasma acidophilum* (SCOP c.66.1.13) with the following scores: probability of 97.05 and E-value of 1.5e-06 for a 110 columns profile-profile alignment (see figure 30 for details). As we did with the first example, we proceeded to structurally superimpose the homologous regions in both folds detected by sequence. We generated an alignment of 118 C-alphas with a RMSD of 3.0 Å. The Z-score was 6.67.

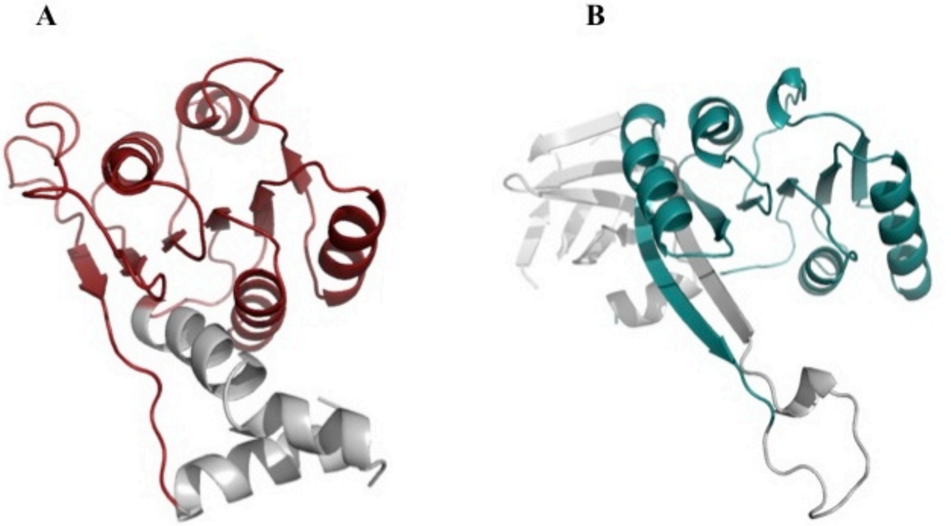


Fig. 30: Two three-layered folds share a conserved fragment. The NAD(P)-binding Rossmann-fold (A) shares a sub-domain size fragment with the S-adenosyl-L-methionine-dependent methyltransferase fold (B).

Among the results we detected an especially remarkable case of hit between two different SCOP fold classes. The DNA-binding protein Sso10a from *Sulfolobus solfataricus* (SCOP a.4) that belongs to the all-alpha fold DNA/RNA-binding 3-helical bundle found with very high probability the $(\alpha\beta)_8$-barrel HemN from *Escherichia coli*. The alignment consisted of 54 columns, with HHsearch probability of 90.36 and reported E-value of 0.02. At the first glance, this hit would look like a false positive because the aligned pairs are classified in different classes. Pairs form different classes would be by definition not easy to align because of the topology of their secondary structural elements. However, the structural alignment returned very good scores: 36 c-alpha carbons (RMSD of 2.2 A) and Z-score of 6.0.

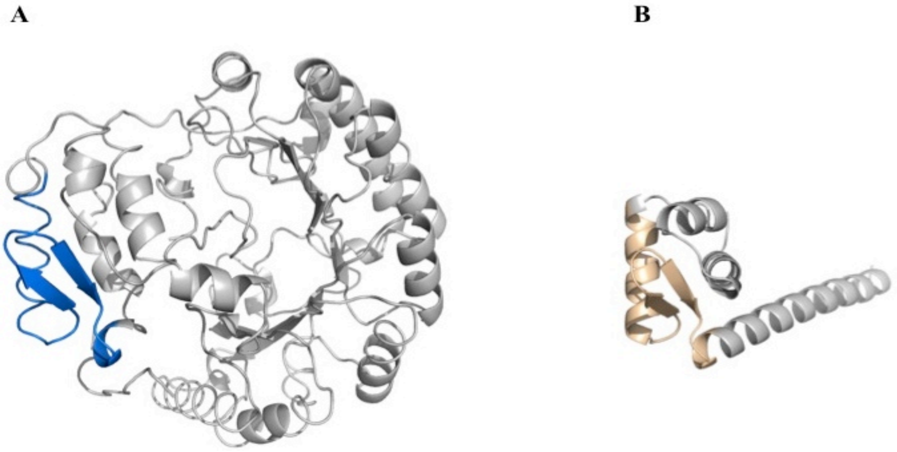


Fig. 31: A super-secondary structure is shared between members of different SCOP classes. The $(\alpha\beta)_8$-barrel 1olt (A) shares an $\alpha/\beta/\beta$ element with the alpha helical bundle 1r7j (B).

The structural comparison depicted in figure 31 can be interpreted in several ways. It can be seen as a false positive found by the *HHsearch* algorithm. On the other hand, it can reflect the subjective nature of the SCOP classification. The aligned fragment between the barrel and the helical bundle is not part of the canonical barrel fold. The probability is among the highest ranked probability hits, it could therefore imply that this fragment is indeed an ancestral reminiscent fossil in both folds.

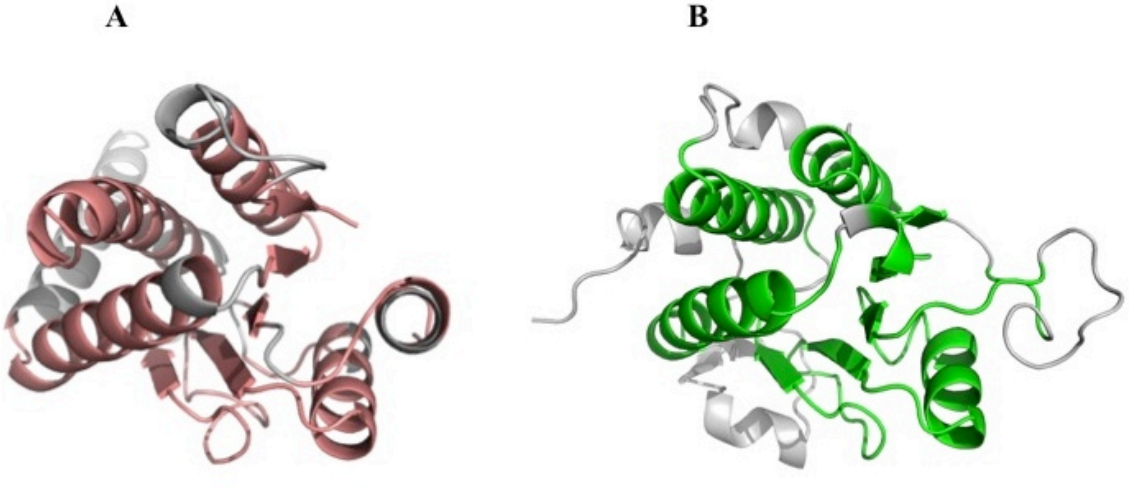


Fig. 32: The SIS fold (A) shares the typical three-layer architecture of the flavodoxin-like fold (B).

In figure 32 we found a shared fragment between the ribosomal protein S2 from *Archaeoglobus fulgidus* (SCOP c.23.15.1, flavodoxin-like fold) and the phosphoheptose isomerase GmhA1 from *Vibrio cholerae* (SCOP c.80.1.3, SIS domain fold). The profile-profile alignment between the sequences of 1X94 (protein S2) and 1VI6 (phosphoheptose isomerase) returned the following values: 85.48 HHsearch probability, 133 columns aligned and E-value 0.11. The structural alignment returned the following values: 121 C-alpha carbons aligned, RMDS 2.3 and Z-score of 6.7. This superposition highlighted a very long sub-domain size fragment conserved between both folds.

It was clear that a systematic exploration of the most ancestral folds (table 18) might provide useful information on early fold evolution. Especially if we can develop evolutionary models that can explain how the most basal folds gave rise to modern ones. Along these lines, the profile-profile comparisons were primarily envisioned as a methodology to find sub-domain size fragments between different folds that could be combined to design well-folded proteins with native-like properties.

The pairwise relationships between different folds had to be automatically explored. Therefore, we started to work in collaboration with Saacnicteh Toledo and Matthias Schwer. They developed an algorithm capable of filtering the results of the profile-profile comparisons in function of multiple features: number of aligned columns, probability score, number of gaps, P-value, E-value. Moreover, an extension of their code performs multiple and dynamic structural superpositions to visualize shared fragments. This algorithm allowed us to discover three folds linked with high probability: $(\alpha\beta)_8$-barrel (SCOP c.1), flavodoxin-like (SCOP c.23) and the Periplasmic binding protein-like I fold (SCOP c.93). In the next chapter, we discuss the construction of a chimeric protein using fragments from Periplasmic binding protein-like I and flavodoxin-like fold structures.

3.12 Mimicking fold evolution by combination of homologous fold fragments.

In the last chapter we described a triple high probability connection between three folds: $(\alpha\beta)_8$-barrel (SCOP c.1), flavodoxin-like (SCOP c.23) and the PBP-like I (SCOP c.93). The algorithm for data mining generated by our co-workers allowed this finding.

We looked for suitable parental template proteins to be combined into a well folded-chimera. In order to decide which proteins select as starting scaffolds we took into consideration many points: high probability profile-profile pairwise alignment, thermo-stability, single domain existence in databases, how easy can we handle the protein in the lab, etc.

The first parental scaffold was CheY from *Thermotoga maritima* (SCOP c.23.1.1, flavodoxin-like fold) because we had experience handling this protein in the lab; it is thermo-stable and can be solubly expressed as a single domain. The second scaffold was the leucine-binding protein (LBP) from *Escherichia coli* (SCOP c.93.1.1, Periplasmic binding protein-like I fold) because it was also easy to handle, there were crystal structures in different conformations available in the PDB and because our general interest in this scaffold for protein design.

3.13 Sequence-based alignments are used to structurally align the flavodoxin-like and PBP-like I folds.

In order to build the chimera LBP-CheY we used the sequences of both proteins to perform a pairwise profile-profile search of the SCOP database. We searched the SCOP95 (SCOP filtered at 95% sequence identity) with default parameters (turning off the secondary structure scoring), to double check the homologous relationship between the flavodoxin-like fold and the periplasmic binding protein. We started the bidirectional search hoping to converge to the same fragment in both directions. The likelihood of having a true positive hit is higher if both searches converge to the same hit with equivalent probabilities.

A pairwise profile-profile alignment between the sequences of 1USG and 1U0Y converged to the same aligned region in the longer protein, 1USG (see figure 33 for details).

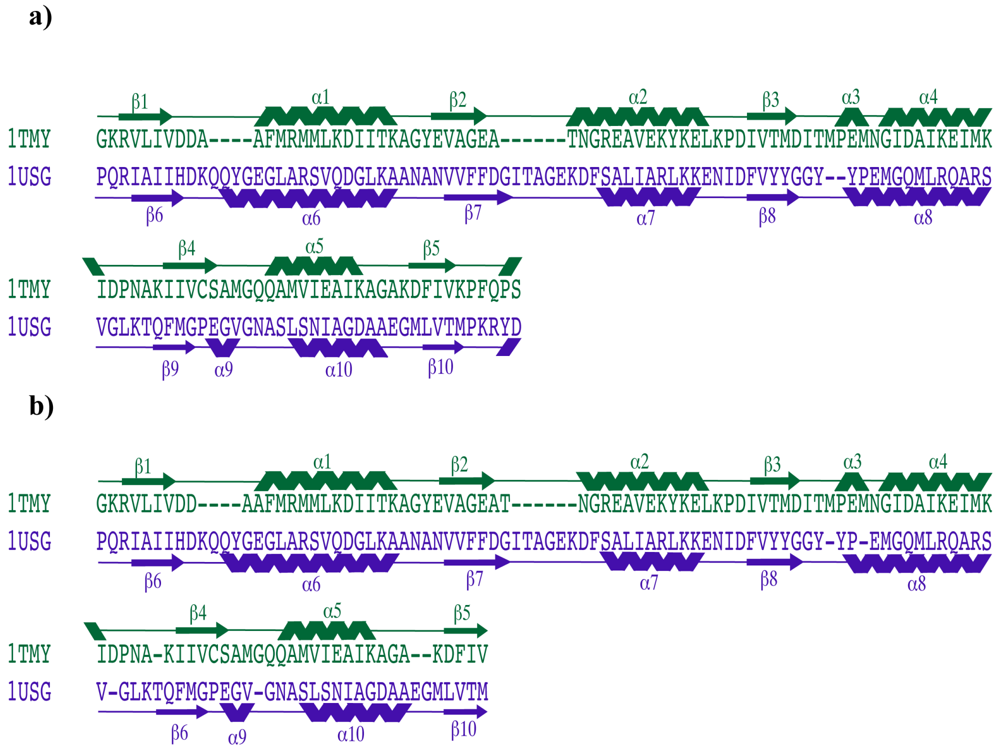

a)
β1
α1
β2
α2
β3
α3
α4
1TMY GKRVLIVDDA----AFMRMMLKDIITKAGYEVAGEA------TNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMK
1USG PQRIAIIHDKQQYGEGLARSVQDGLKAANANVVFFDGITAGEKDFSALIARLKKENIDFVYYGGY--YPEMGQMLRQARS
β6
α6
β7
α7
β8
α8
β4
α5
β5
1TMY IDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVKPFQPS
1USG VGLKTQFMGPEGVGNASLSNIAGDAAEGMLVTMPKRYD
β9
α9
α10
β10
b)
β1
α1
β2
α2
β3
α3
α4
1TMY GKRVLIVDD----AAFMRMMLKDIITKAGYEVAGEAT------NGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMK
1USG PQRIAIIHDKQQYGEGLARSVQDGLKAANANVVFFDGITAGEKDFSALIARLKKENIDFVYYGGY-YP-EMGQMLRQARS
β6
α6
β7
α7
β8
α8
β4
α5
β5
1TMY IDPNA-KIIVCSAMGQQAMVIEAIKAGA--KDFIV
1USG V-GLKTQFMGPEGV-GNASLSNIAGDAAEGMLVTM
β6
α9
α10
β10

Fig. 33: The sequence (**a**) and structure (**b**) based alignments of 1TMY and 1USG. The scores for the sequence alignment are as follows: a) probability of 81.19, b) identities=15% and c) the number of aligned residues was 106. The structural scores are as follows: an R.M.S.D of 2.667 **Å** for 95 c-α atoms with a Z-score of 5.35.

Two intertwined lobes compose the periplasmic binding protein (figure 34). The N terminal lobe crosses the hinge region to the C terminal lobe at the end of β5. The β-sheet from the C-terminal lobe starts in β6 and goes back to the N-terminal lobe in β10, to finally cross one more time after β10 to complete the extended β-sheet from lobe C with a β-meander (composed by the last two β-strands). The profile-profile pairwise alignment covers the CheY fold sequence from residue 2 until residue 109 from β1 to the beggining of α5. In LBP, the alignment goes from residue 138 starting in β6 finishing in residue 253 in β10 (see figure 34).

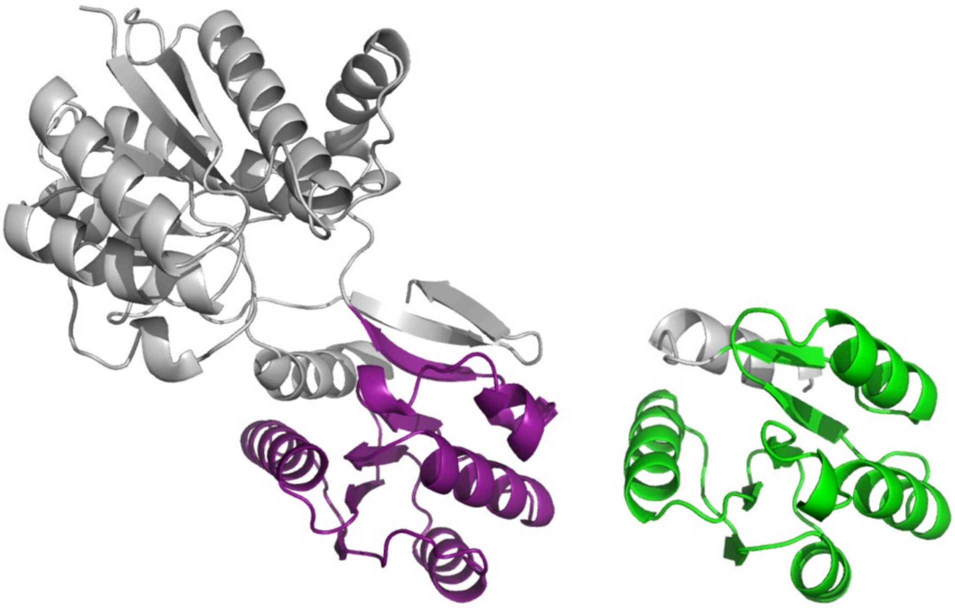


Fig. 34: Structural superposition between 1USG (violet region) and 1TMY (green region). The alignment showed a conserved sub-domain size fragment between both folds. We generated a chimeric domain by replacing the violet region with the green fragment.

We used the profile-profile alignment from figure 33 to guide a structural superposition of the two proteins shown in figure 34. The web server SSM allows users to perform range-restricted structural alignments; we therefore structurally aligned the regions of the folds that were aligned by HHsearch.

The structural superposition produced the following scores: rmsd of 2.667 Å over 95 Cα atoms and Z-score of 5.354. Interestingly, the Z-score reported for this alignment lies within the so-called twilight zone of structural homology: Z-score values below 2 are considered insignificant, between 2-8 a gray zone of homology is determined and above 8 the pairs aligned are considered homologous proteins. The structural alignment with Z-score of 5.3 lies directly in between 2 and 8 therefore it is not possible to determine, with this single score, whether these two folds are homologous. Nonetheless, taking into account the bidirectional pairwise profile-profile alignment between LBP and CheY we can conclude that both folds are at least locally homologous.

Guided by the previous alignments (structural and sequence-based) we defined the region that would be interchanged between both folds. We assumed that the flavodoxin-like fold is homologous to the aligned region on the second C-terminal lobe of the leucine-binding protein. The alignments cover as much as three αβ elements plus an extra β-strand. We intended to interchange these elements to create chimeric proteins that would resemble how evolution would have created structural diversity starting from a set of reduced fragments.

3.14 Homology modeling and computational chimera design

We produced several versions of the LBP-CheY chimera. Versions 01 and 02 were initially tested for native like folding properties. Versions 03, 04 and 05 were attempts to generate protein crystals that would allow the determination of an X-ray crystal structure. See appendix (section 8.2) and figure 35 for details about the different versions of LPB-CheY.

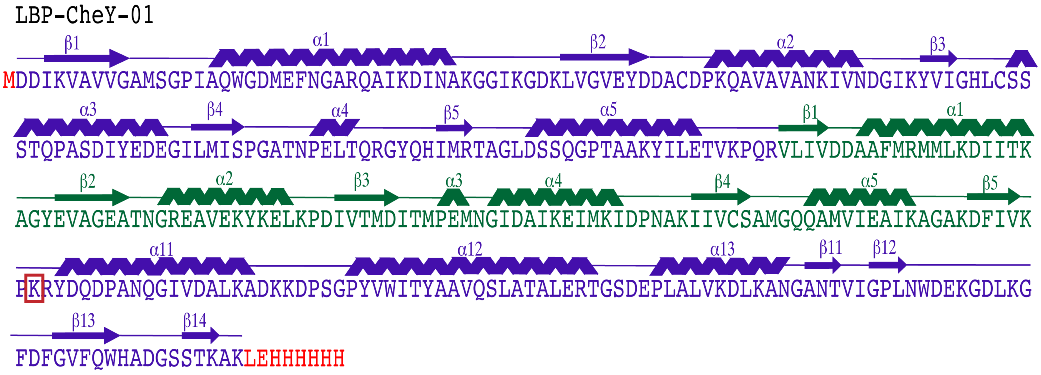

LBP-CheY-01
β1
α1
β2
α2
β3
MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSS
α3
β4
α4
β5
α5
β1
α1
STQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQRVLIVDDAAFMRMMLKDIITK
β2
α2
β3
α3
α4
β4
α5
β5
AGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVK
α11
α12
α13
β11
β12
PKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKG
β13
β14
FDFGVFQWHADGSSTKAKLEHHHHHH

Fig. 35: LBP-CheY-01 secondary structure representation. In the design procedure we did not include a lysine (highlighted in red). We engineered an N-terminal methionine and a C-terminal 6X-His tag. We introduced the missing lysine in LBP-CheY-02 (see section 8.2 from appendix for sequence details on the different versions of the LBP-CheY chimera)

For the chimera building process we generated several homology models of all the chimeras. It is difficult to predict how the pieces from different folds are going to interact with each other; however, by using a comparative modeling protocol we could estimate the probable tridimensional arrangement. The homology modeling process involves four steps: fold assignment, target-template alignment, model building and evaluation of the model. We used as templates the coordinates from the leucine-binding protein from *Escherichia coli* (1USG) and the response regulator CheY (1TMY) from *Thermotoga maritima*.

We generated an alignment using the sequences from both folds in PIR format, see appendix for details (section 8.1). The alignment between the sequences of both structures is crucial because it will determine the transition of one sub-domain size fragment to the next one. We aligned the HMM of both structures in the HHpred web-server, this kind of alignment is more precise than aligning single sequences.

We checked the homology model using an internal evaluation protocol of Modeller. The evaluation protocol is called ANOLEA, where non-local interactions are used to assess the quality of the model. Of course, the most obvious assessment of the model is the one made by performing a careful visual inspection. A well-folded protein will feature a packed core with hydrophobic residues shielded from solvent.

Modeller performs a raw minimization protocol. To evaluate the models more precisely we employed energy minimization with the program Rosetta. The idealize protocol was used, because it restructures a protein molecule by adjusting the bond lengths, and bond and torsion angles to idealized values. This protocol can yield structures with bad clashes; therefore, we subsequently used the *rosetta-relax* protocol. This approach is used to lower the energy of a model through minor changes to the backbone and side-chain torsion angles. The Rosetta relax algorithm can generate 100 structures (a higher number of structures can be specified) that will have different *rosetta-energy* units. Rosetta does not calculate physical energies (i.e. kcal/mol); the calculated energy cannot be translated into physical energies but are useful to compare different structures from the same run.

We performed the *Rosetta*-protocol for all chimeric versions. The protocol we developed was useful to spot a missing lysine in LBP-CheY-01 (see figure 35 for details). The energy of the residues around the insertion site (two residues after β4) was unusually high. We observed that the peptide bond-length was quite far from the knowledge-based distribution of peptide-bond lengths in proteins. The missing lysine was added to version LBP-CheY-02. In general, the energy per residue of the different chimeras did not show any difference that could lead us to choose different insertion sites. No chimera consistently displayed a better energy per residue than the rest. In other words, besides spotting errors in the design, we did not find any consistent significant difference in energy between all the versions (versions 1 to 5, shown in appendix section 8.2).

3.15 Experimental characterization of the chimera LBP-CheY

We constructed and tested LBP-CheY-01 and LBP-CheY-02 in the laboratory. No major differences between both chimeras were seen during the biophysical characterization. Therefore we show only the data for LBP-CheY-01.

The chimera LBP-CheY-01 displayed native-like properties and could be purified as a monomer. Nonetheless, we were not able to produce crystals. Therefore we generated several variants of the initial construct. The data generated from the partial characterization of all chimeras are outlined in table 20.

Table 20: Experimental characterization of LBP-CheY chimeras. The amount of protein, in the soluble or insoluble the fraction, is indicated by the number of (✓) symbols. Success or failure of the crystallization attempts is indicated with (✓) or (X) respectively.

	Soluble	**Insoluble**	**Refolding**	**Monomer**	**Multiple oligomerization states**	**Secondary structure content (CD spectra)**	**Tertiary structure content (Fluorescence)**	**Enough yield to set crystal screening**	**Crystallized**
LBP-CheY-01	✓	✓✓	✓	✓	✓	✓	✓	✓	*X*
LBP-CheY-02	✓	✓✓	✓	✓	✓	✓	✓	✓	*X*
LBP-CheY-03	✓	✓✓	✓	✓	✓			✓	7 Å
LBP-CheY-04	✓	✓✓	✓	✓	✓			✓	7 Å
LBP-CheY-05	✓	✓✓	✓	✓	✓			✓	*X*

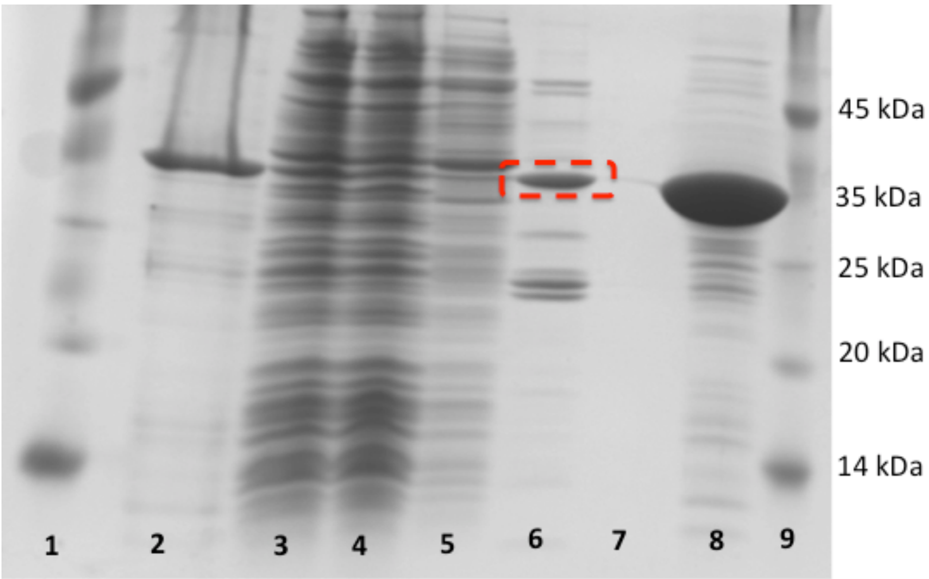


Fig. 36: Nickel affinity purification of LBP-CheY-01. An increasing percentage of imidazole was used for elution. The purification and refolding of LBP-CheY-01 was evaluated by SDS-PAGE. Lanes: 1) molecular weight marker, 2) insoluble fraction, 3) soluble fraction, 4) flow-through, 5) wash 5%, 6) elution 40%, 7) elution 80%, 8) refolded protein and 9) molecular weight marker. A red dashed square indicates the LBP-CheY-01 band at around 35 kDa.

We cloned the chimeras and LBP fused to a 6X His Tag. All the proteins were expressed in *Escherichia coli* using a standard protocol (see methods for details). LBP-CheY-01 was mainly expressed in the insoluble fraction (figure 36); however, some small amount (between 15 to 20 % of the total protein) was still soluble. LBP wild type was mainly expressed in the soluble fraction. Nickel affinity chromatography was the first purification step. A second purification step involved size exclusion chromatography.

We performed analytical gel filtration runs with purified samples to estimate the oligomerization state of the chimera. The gel filtration run showed that it is possible to purify the chimera LBP-CheY-01 as a single monomeric protein and the size is similar to the size of the LBP (figure 37). LBP-CheY-01 eluted later than LBP; this could indicate that LBP wild type is more compact than the chimera.

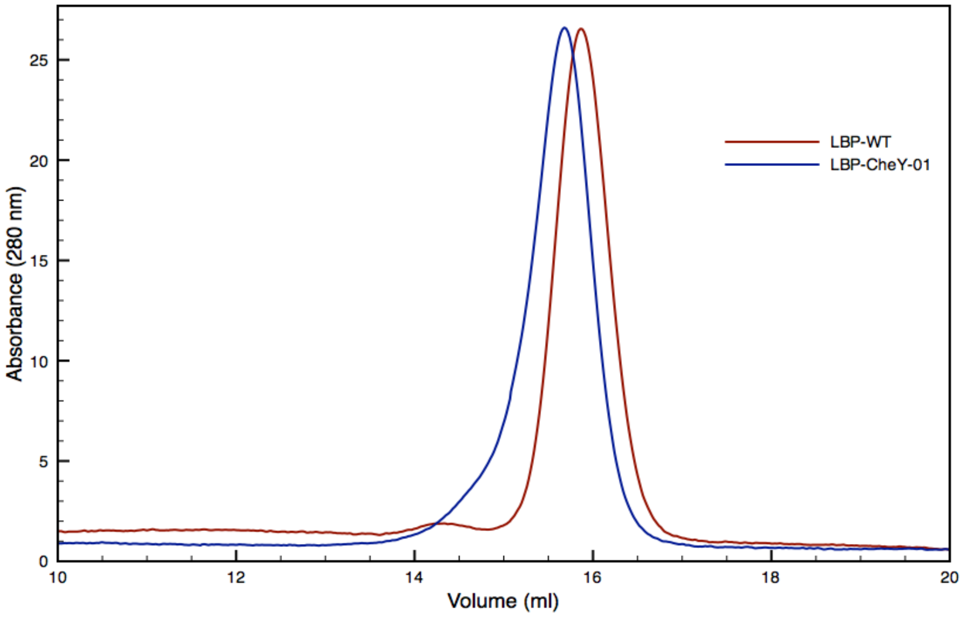


Fig. 37: Analytical gel filtration curves of LBP (red) and LBP-CheY-01 (blue). The calculated apparent molecular sizes of LBP wild type and LBP-CheY-01 were 39.7 and 42.3 kDa respectively.

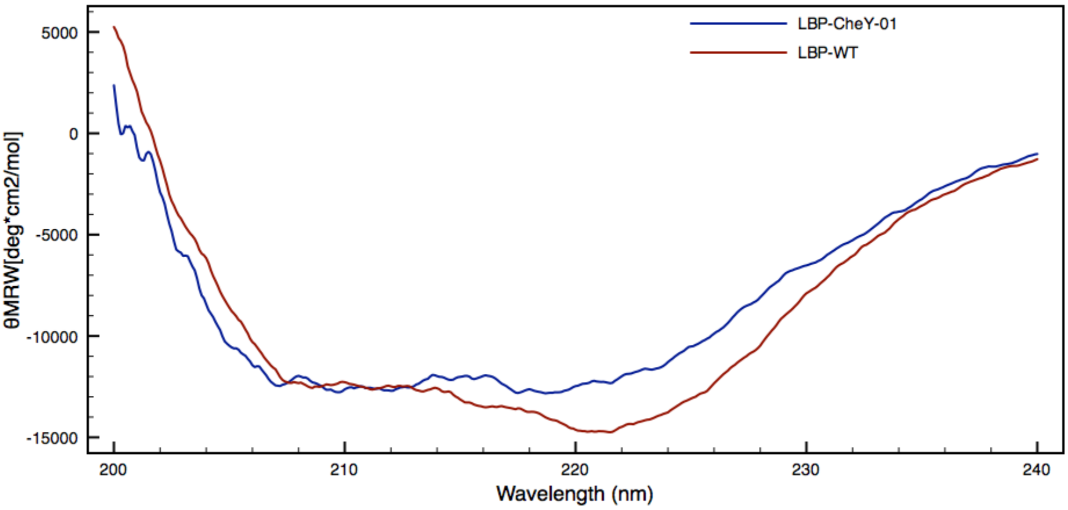


Fig. 38: Circular dichroism of LBP (red curve) and LBP-CheY-01 (blue curve). The secondary structure content of both proteins is similar. Nonetheless, the alpha-helical content is more pronounced in LBP wild type.

We measured far-UV circular dichroism to evaluate the secondary structure content of the chimera in comparison to the wild type LBP. The CD-spectra showed that both proteins have a very similar content of secondary structures. The signal for the alpha helical content is a bit stronger in the wild type protein than in the chimera. Nonetheless, the differences are small (figure 38).

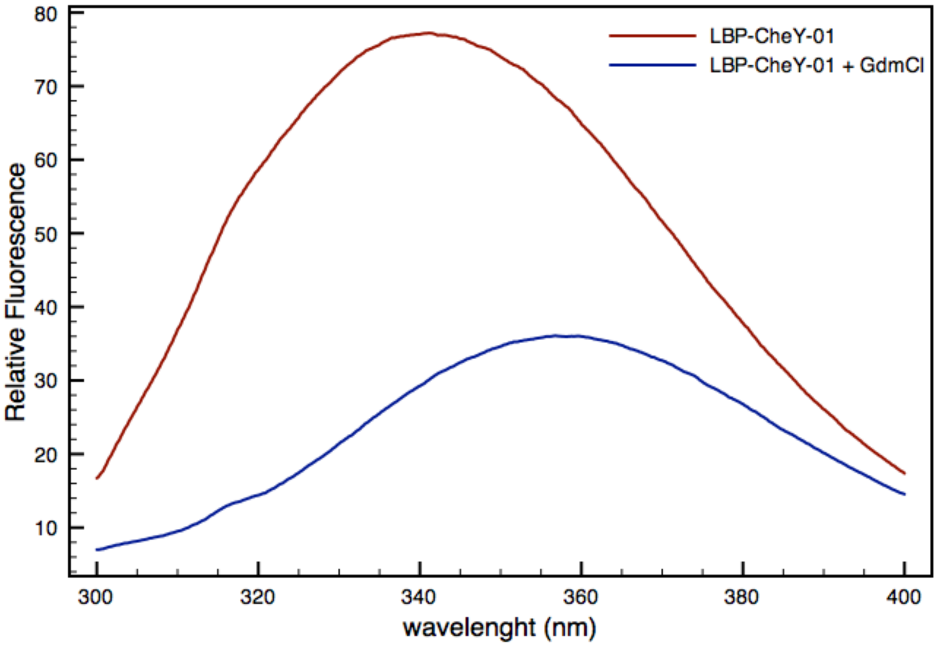


Fig. 39: Fluorescence spectroscopy curve of LBP-CheY-01. In red we observe the natively folded chimera with an emission maximum at 342 nm. After addition of 6 M GdmCl the signal is reduced and shifted.

In order to assess the tertiary structure of the chimera, we performed fluorescence spectroscopy measurements (Figure 39). The protein showed a maximum at around 340 nm indicating that the aromatic residues were shielded from the solvent. To evaluate how the protein unfolds we added 6 Molar Guanidine hydrochloride and we measured the spectrum again. The spectrum showed a maximum at 360 nm and lost about a third of its intensity. This is a clear indication that the protein was previously folded and after the addition of the guanidine hydrochloride, the aromatic residues were exposed to the solvent, causing a shift and a reduction in the relative fluorescence.

3.16 LBP-CheY crystallization efforts

To evaluate how the two fragments from the different folds fit with each other we decided to produce an atomic model of the well-folded chimera. We expressed LBP-CheY-01 and LBP-CheY-02 in large scale to set crystallization screenings because we needed big amounts of protein. Both chimeras were mainly expressed in inclusion bodies; only a few fraction of the total protein was soluble expressed. Therefore, we refolded LBP-CheY-01 and LBP-CheY-02 from the insoluble fraction. Generally, the protocol yielded 20 mg of pure protein starting from 4 liters of culture media. The success of the refolding protocol was evaluated by SDS-PAGE (lane 8 in figure 36).

The chimeras were refolded into different oligomerization states. For crystallization purposes it was necessary to produce a monodisperse sample; therefore, we implemented a two-step purification protocol after the refolding to obtain a sample for crystallization. First, we used ion exchange to enrich the monomeric state over other states in our samples. Finally, we used size exclusion to isolate only the monomeric version of the chimeras. We concentrated samples of LBP-CheY-01 and LBP-CheY-02 up to 10.0 mg/mL (400 µL each sample) to set a crystallization screening (6 plates x 96 conditions). Unfortunately no crystals were produced from these crystallization attempts.

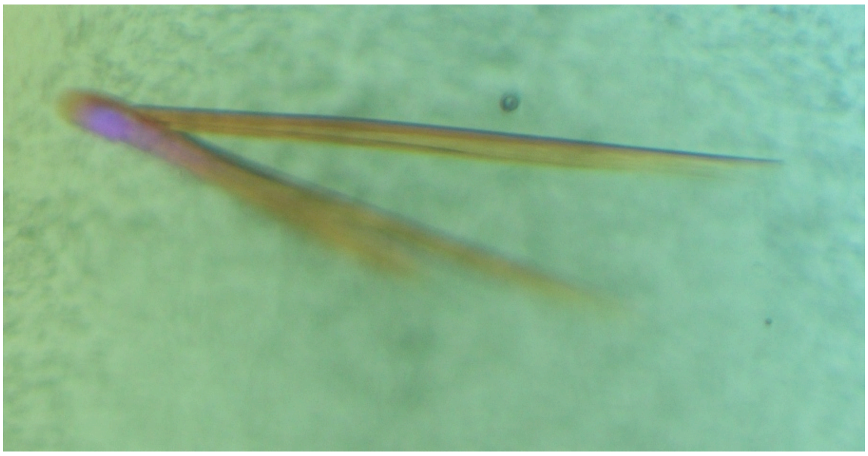


Fig. 40: Crystals of LBP-CheY-03. The protein concentration was 23 mg/mL, and the conditions were: 0.2 M Ca Acetate, 0.1 M Na cacodylate pH 6.5 and 18% PEG 8000 (w/v). These protein crystals diffracted up to 9.0 Å resolution.

We judged that the flexibility of the 6X His Tag represented a problem for crystallization. Moreover, our purification protocol did not require the presence of such a tag. Therefore, we removed it to produce chimera LBP-CheY-03. We followed our established purification protocol from inclusion bodies to obtain a monodisperse sample that was concentrated up to 23 mg/ml. We set a crystallization screening using two drops, one with 20-fold molar excess of leucine and another drop only with protein. There was a crystal hit for the drop without leucine [mother liquor components: 0.2 M Ca Acetate, 0.1 M Na cacodylate pH 6.5 and 18% PEG 8000 (w/v)]. The crystals were needle-like shaped and grew on the top of each other (figure 40). We shot these crystals at the PXII beamline at the Swiss light source to confirm that they were protein crystals. The crystals diffracted with maximum resolution of 9.0 Å.

Visually, these crystals looked needle-like shaped with sharp ends and had a considerable size. We set optimization trials around the original conditions to subsequently fish and screen around 40 crystals. The variations only included the concentration of the contents of the drop and no additives. Unfortunately the best diffracting crystal only reach 7 Å. Interestingly, and as expected, the improvement in the diffraction followed an increase in the concentration of the precipitant, from 18 to 22% PEG 8000 (w/v).

Subsequently, we tried several documented optimization techniques to improve the diffraction of the crystals: a) dehydration of the crystal directly at the beamline, b) dehydration of the crystal in the mother solution drop by direct contact with air, c) subsequent removal of solvent by transferring crystals from mother solutions with increasing precipitant. Yet, we did not manage to improve the crystals to reach a resolution that allowed structure determination. We then proceeded to re-design the LBP-CheY chimera in two subsequent experiments to create LBP-CheY-04 and LBP-CheY-05.

LBP-CheY-04 is a design suggested by an unpublished algorithm (Nils Woetzel, personal communication) named BCL::FusionProtein. This algorithm takes into account several scores (e.g. clashes introduced, peptide bond distance between the fusion fragments) to suggest an insertion site in the acceptor scaffold. The algorithm localizes the insertion sites by cutting and pasting the donor fragment in the secondary structural elements.

The difference between LBP-CheY 3 and 4 are the insertion points in the acceptor scaffold: chimera LBP-CheY-03 has the flavodoxin-like fold structure (starting at β1) inserted right at the end of the loop α5β6. In contrast, LBP-CheY-04 features the insertion of CheY after the first residue of β6 so we have a final fusion of β6 from LBP to β1 from CheY. The c-terminal insertion point is the fusion between β5 from CheY and β10 from LBP, right in the middle of the β-strand. This kind of configuration may yield a more rigid construct.

We cloned, expressed and purified LBP-CheY-04 following the same protocol used for the other chimeras. During the purification process the chimera behaved in a similar manner than previous constructs. We did not perform a full biophysical characterization. A crystallization screening was set with the purified sample. From this attempt, we again produced good-looking crystals that diffracted no better than 7.0 Å resolution. Then we decided to redesign the construct to produce LBP-CheY-05 in an attempt to improve the resolution of the crystals.

3.17 LBP-CheY loop re-design

For the design of chimeras 1 to 4 we used as template the open conformation of the leucine-binding Protein (1USG). For chimera 05 we switched our template to 1USK, that is the closed conformation of LBP. The region we estimated that might prevent compact packing in the chimera is highlighted in figure 41; the superposition shows that the loop β4α5 in CheY is clearly bigger than the topologically equivalent loop in LBP (booth loops crossed by red bars). Moreover, the loop from CheY is not only abolishing interactions for binding leucine but also it may be clashing with the upper lobe of LBP. The composition of the loops is very different in the two proteins: ***CSAMGQ*** in CheY and ***GPEGVGN*** (comprising a 3_{10} helix) loop in LBP.

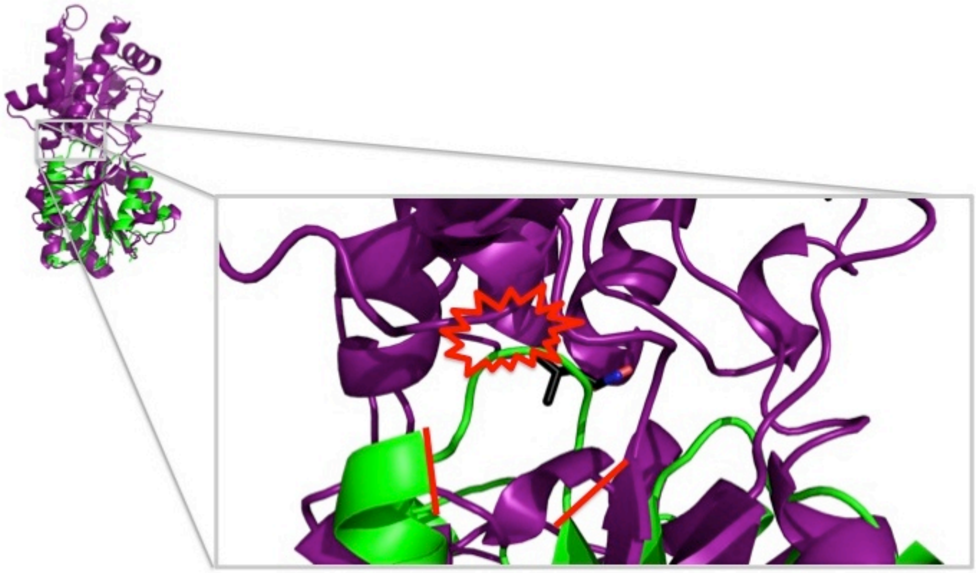


Fig. 41: Zoom in into the superposition of 1TMY (green) and 1USK(violet). Two red bars delimit the loops that are different in length. In LBP, the loop includes 3_{10}-helix. A red star indicates a probable clash in the chimeric protein that is preventing a close conformation. Leucine is highlighted in black sticks.

Observing closely the binding pocket of 1USK we discovered that 9 residues establish direct contact with leucine. See table 21 for details.

Table 21: Residues in 1USK that do not have an equivalent in the chimera LBP-CheY. Also we highlight whether the residues would be clashing in the new chimeric context.

Residue in LBP (1USK)	Presence/absence (by replacing with loop from CheY)
S79	Present
G100	Present
T102	Present
Y202	Missing
E226	Missing
W18 (upper lobe)	Present in LBP but maybe clashing with CheY loop
Y150	Missing
G227	Missing
Y276 (upper lobe)	Present in LBP but maybe clashing with CheY loop

In LBP-CheY-03 two interactions with the ligand are missing, the ones established by E226 and G227. We think these residues are important to maintain the geometry of the binding pocket. What is more, two other interactions established by W18 and Y276 with the hydrophobic lateral chain of leucine, are also probably abolished. It appears that the CheY loop is clashing with the hydrophobic residues W18 and Y276 (see figure 41 for details).

In order to build a design capable to adopt a closed conformation, that in principle could produce better crystals, we built chimera 05. In this chimera, the introduced CheY fragment is shorter than in previous chimeras; it includes up to the end of β4 of CheY. We therefore hoped to recover many interactions abolished in previous chimeras. We cloned, expressed and purified LBP-CheY-05 as we did with previous versions. We also set a crystallization screening with this version. Unfortunately, no crystals were produced.

4. Discussion

4.1 Profile-profile comparisons of two superfolds

As Charles Darwin recognized similarities between different species, protein scientists observe sequence, structure and functional shared patterns among proteins (*10*) (*73*). Through these observations, general sets of rules governing protein folding and function start to emerge. For instance, statistical studies of protein sequences revealed independent networks of coevolving functional residues that are more sensitive to mutation than others (*74*). Furthermore, comparative structural studies have shown how isolated protein domains with primitive functions can evolve to yield multi-domain proteins with sophisticated activities (*75*). Through the recognition of energetically unfavorable features in natural proteins, which arose by natural selection and random genetic drift, the design of ideal protein scaffolds (*76*) was successfully accomplished. These findings are the natural result of our better understanding of protein evolution.

Our understanding of protein evolution is in constant expansion. Not only the development of novel tools for homology detection contribute to this advancement. Also protein-engineering experiments test our knowledge of the rules governing protein folding and function. With this work, we identified homology between proteins that adopt different folds. Structure-based evidences (*36*) and protein engineering experiments (*37, 40*) previously suggested this possibility. We then decided to use the best available sequence-based homology detection tools, working with the SCOP classification system, to explore the structures grouped into these folds.

The exploration of the SCOP superfamilies from different folds by pairwise alignment of HMMs provided initial evidences for homology. This recently developed tool for homology detection is already used by many of the best structure prediction servers and has ranked best in last years CASP exercises (*77*). While the algorithm has proven to be one of the best of its class, at the time of its first benchmarking there were many hits between different folds, which could not be easily explained (*33*).

The high sensitivity of the HHsearch tool inspired sequence-based explorations of the structure space. For instance, by exploring the SCOP20 database (sequences are similar to a maximum of 20% sequence identity) Alva et al. presented a starting point in which a single image gives an overview of how some representative members of protein families and superfamilies cluster based on sequence. They showed that many of the superfamilies of one fold cluster together and that there appears to be more connectivity in fold space than previously expected.

We then recognized the potential of HHsearch combined with structure-based observations to assess probable homologous relationships between distinct fold classes. As starting point of our research, we re-evaluated the sequence-based connections that support homology among the $(\beta\alpha)_8$-barrel fold superfamilies. We indeed reproduced previous results and extended it by connecting up to 30 from 33 superfamilies. The most connected $(\beta\alpha)_8$-barrel fold proteins are the aldolases; not only were strongly connected to many phosphate binding barrels, but also with the two flavodoxin-like fold superfamilies that share local structural similarity to $(\beta\alpha)_8$-barrel fold proteins. The 3 $(\alpha\beta)_8$-barrel superfamilies that do not hit each other with more than 80% probability may not be well represented in the structural databases. Perhaps, as more structures, or sequences, become available, it will be possible to link all $(\beta\alpha)_8$-barrel fold superfamilies. On the flavodoxin-like fold side, the connectivity was less pronounced. Lacking enough structures in the databases, as previously discussed with the barrels, may be the reason of such observation. On the other hand, three-layered architectures, with variable topological connections, are rather common among many proteins in the databases. In consequence, a convergent origin of this fold seems plausible.

The most interesting part comes when we evaluated the inter-fold connections. Strikingly, we found many connections between two flavodoxin-like superfamilies, the B12-binding domains and response regulators, with a number of $(\beta\alpha)_8$-barrel fold superfamilies. We had to be cautious at this point, because every computational algorithm would in principle deliver a number of false positives.

Having this in mind, we set astringent measures to reduce the number of false positive instances: a) the probability connections had to be reciprocal, b) a stringent cut-off of 80% Bayesian probability, c) secondary structure content was not considered, and d) many proteins from the same superfamily had to hit many members of the other superfamily-fold.

Figure 8 shows the density of hits in function of probability by starting the HHsearch job with either a flavodoxin-like or a $(\beta\alpha)_8$-barrel fold query. Here it is possible to appreciate that a high density of hits from the same fold class (self-hits) are indeed found at higher probabilities than 80%. What is more, starting the search with $(\beta\alpha)_8$-barrel fold queries, we found the same number of flavodoxin-like hits than $(\beta\alpha)_8$-barrel fold hits at around 80% probability.

This overlap in the number of hits from both fold classes emphasizes the strong sequence-based connectivity. The flavodoxin-like fold queries found other flavodoxin-like fold hits in a very narrow probability space. Nevertheless, at very high probability many members of the $(\beta\alpha)_8$-barrel fold are found, at least two times more members than any other fold. One may argue that the secondary structure content, or the similar architecture of the αβ elements could be the source of these high-scored relationships. We can rule out this possibility since only two from 15 members of the flavodoxin-like fold were hitting with such high probability scores a number of $(\beta\alpha)_8$-barrel fold superfamilies.

4.2 Searching the databases for intermediate sequences.

The profile-profile searches allowed us to locate the sequence space interface between both folds where it was most probable to discover sequences with intermediate features. We launched remote homologous searches using queries from both folds that were closely related. The searches were a relatively straightforward approach; the most difficult part of our research was the clustering and evaluation process of the intermediate sequences. Several factors made the evaluation a daunting task:

1. The sequences coming from the raw HHsenser outputs are shorter than the complete hit, because the high scoring pair of the BLAST comparison is restricted to a local region of the query-hit alignment. In principle we could have extracted the full-length sequences found by HHsenser and cluster full domain proteins; the complication comes when the domain boundary has to be allocated. This is a very well known problem in protein evolution studies because domain boundary definitions are neither trivial nor easy. It usually requires the building of a single multiple sequence alignment per hit, and we clustered thousands of sequences. We could have clustered the sequences by families to well align them separately in order to define the domain boundaries.
2. These sequences are not only short; they have missing residues as well. HHsenser only keeps the high scoring pair that matches hit and query, it does not keep the rest after seed extension.
3. One sequence can be located in the middle of two clusters (while connecting them), but it could be a false positive. This sequence may be a very short fragment that has similarities to both folds but only locally. As soon as one pulls the complete sequence this hit would not appear in the middle anymore, it would be closer to the members of the cluster that are more similar to it (one or the other fold).
4. A sequence is not in the middle but is connected to clusters from different folds. There is a short fragment-sequence, inside the cluster from fold ***x***, that had strong similarities with one protein in the cluster from fold ***y***.

In this sense, the repulsion and attraction values (selected by default in CLANS) are placing this protein inside of its "own" cluster, even if it had a high scoring pair (represented by edges connecting two clusters) in the cluster from the other fold. Because there are so many sequences in its family-cluster the attraction will be too strong and therefore this sequence would not appear in the middle.

5. Connectivity among the clusters. The connectivity changes dramatically if the P-value of the connections is slightly changed. At a given P-value, a protein may seem a link between two clusters; however, when the complete protein replaces this hit the link may not hold. These borderline connections, can be either established or removed by varying the P-value one order of magnitude.
6. Singletons (or unrelated proteins) pulled in the HHsenser search because very relaxed parameters were used. This would introduce visual noise into the cluster map.

Having all these complications in mind, I looked for other features in the CLANS maps: the position, the connections, the length of the hit, and the P-value of the connection.

We used the sequence of the glutamate mutase (SCOP c.23.6) from *Clostridium cochlearium* to perform the deepest homologous searches in this work. Remarkably, this search travelled the sequence space of the flavodoxin-like fold to find sequences associated with the $(\beta\alpha)_8$-barrel fold. There were several short $(\beta/\alpha)_8$-barrels that have high sequence identity towards both folds. The cluster map of the hits found is shown in figure 42. Highlighted in yellow, we found the sequences that belong to the "B12-binding domain_like associated with radical SAM domain". We also see some sequences that would hypothetically fold as $(\beta/\alpha)_8$-barrels; these are very similar to the sequence of one structure, also present in the map (colored in blue), the nicotinate phosphoribosyltransferase from *Thermoplasma acidophilum* (PBD 1YTD; SCOP c.1.17)

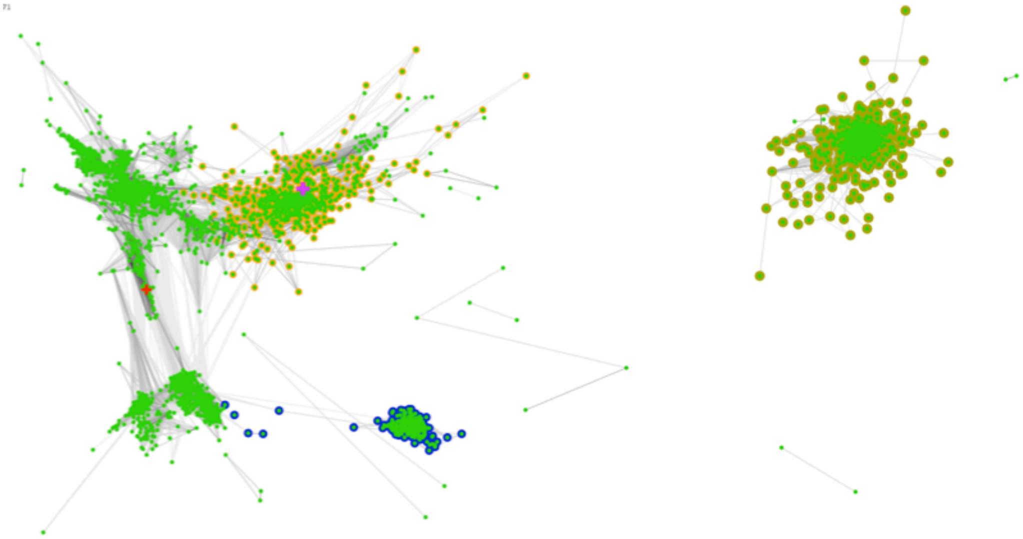


Fig. 42: Clustering map of the HHsenser output from 1I9C (SCOP c.23.6). The coloring is as follows: a) flavodoxin-like fold hits highlighted in green, b) response regulators (CheY like superfamily C.23.1) in green/brown, c) B12_binding_like in green/orange, d) fragments of $(\beta/\alpha)_8$-barrels (in green/blue); among these sequences we can find the sequence from the structure 1YTD (c.1.17.1) and proteins that belong to the $(\beta/\alpha)_8$-barrel IMPDH superfamily (c.1.5.1), e) starting query 1I9C highlighted as red cross and f) NTM0182, highlighted as a violet cross.

Remarkably, it seems that the deep remote homologous search, clustered in figure 42, travelled between the two different structural islands by finding fragments of $(\beta/\alpha)_8$-barrels that are locally similar to both folds. Analyzing these sequences we did not find any kind of intermediate form (in size terms) between the clusters from the different folds (green and blue/green). It seems that there is a very discrete change, in function of the sequence, from one fold to the other. Our rationale was that a single sequence connecting clusters from two folds would be a rarity. Unless the protein is a pseudo-gene, a very special case of non-homologous recombination or its primary sequence evolved far away from any other protein in the database, we would always expect a family of proteins linking the two folds.

PC04 intermediate sequence is a particularly special case since this protein is particularly different from any other protein in its family. A usual BLAST search finds hits that are sorted by the sequence identity respective to the query. The first hit would be 98% identical to the query; the second 90% the tenth hit may be 70% and so on.

With PC04, the first hit is only 45% identical to the query; it seems that the sequence diverged quite a lot. We are still working on getting more structural information on this protein.

In the clustering map from figure 15, we observe the B12_like associated with SAM radical family of proteins that includes NTM0182. It is located closer to the barrel fold group of proteins. This also can be seen in figure 12, which shows a clustering map of 1I9C (c.23.6 B12-binding domain) and 1ZFJ (c.1.5.1: Inosine monophosphate dehydrogenase) with NTM0182 highlighted with an arrow. It is clear that B12_like associated with SAM radical family of proteins is indeed clustered closer to $(\beta/\alpha)_8$-barrels. NTM0182 was chosen for experimental characterization because it was a thermo-stable member of this intermediate family of proteins. We will discuss the structure of NTM0182 in the next section.

Recognizing that CLANS clustering of HHsenser outputs seem to have many limitations, I implemented another early idea. Ideally, an intermediate candidate that would link two protein structures from different folds may be found by queries from both folds. Therefore, it may be present in two HHsenser outputs. I looked for common gene identifiers in flavodoxin-like and $(\beta/\alpha)_8$-barrel HHsenser outputs. However, I never found any common gene identifier.

This idea never really worked out as expected, I assume that hitting the same protein starting with two structures from different folds was very unlikely. Now, after having assessed the B12-binding domain_like associated with radical SAM domain family of proteins is intermediate, I wanted to see whether both folds hit this family.

The B12-binding domain_like family of proteins was clearly more similar to the flavodoxin-like fold than to the $(\beta/\alpha)_8$-barrel. Therefore, it appears in many of the flavodoxin-like fold HHsenser outputs (specially the ones that were made using queries from the superfamily c.23.6). Thus, finding the same family of proteins in $(\beta/\alpha)_8$-barrel HHsenser outputs would be more interesting.

In the HHsenser outputs from the *Methanococcus jannaschii* synthase (SCOP c.1.27.1) and the glycerol uptake operon antiterminator protein from *Thermotoga maritima* (SCOP c.1.29.1) we found the cobalamin B12-binding:Radical SAM from *Geobacter sp*. Moreover, in the output from the dihydropteroate synthetase from staphylococcus aureus (SCOP c.1.21.1) the radical B12-binding:Radical SAM protein from *Methanosaeta thermophile* is also found.

The superfamilies c.1.27.1, c.1.29.1 and c.1.21.1 are hitting flavodoxin-like fold superfamilies with high probabilities (around 80%). The presence of the B12-binding domain_like associated with radical SAM domain proteins in both, the flavodoxin-like and $(\beta/\alpha)_8$-barrel fold HHsenser outputs confirms my assumptions. This family is a convergence point for both folds. For future work, we will build clustering maps using the HHsenser outputs from both folds where the B12-binding domain_like associated with radical SAM family was found.

4.3 Lessons learned from NTM0182 structure

The determination of the NTM0182 structure provided a valuable resource of information to better understand the evolution of the $(\beta/\alpha)_8$-barrel and flavodoxin-like folds. However, from the pure structural analysis, it is not clear yet whether the β-strand swap has any biological significance or not. The β-strand invasion may be a consequence of the construct: without the C-terminal domain partner, the binding of the wild type ligand may be abolished. In consequence, the β-strand swaps to shield a free hydrophobic pocket. We hypothesized that the wild-type ligand may have a phosphate moiety. From some preliminary analysis, performed with the full-length construct that includes the C-terminal SAM radical domain, we observe a very strong stabilization effect of the protein after dialyzing with phosphate buffer. More experiments are needed: test stabilization of NTM0182 (monomer and dimer) upon phosphate addition and estimation of the oligomeric arrangement of the full-length construct.

By comparing the NTM0182 structure with members of the (β/α)$_8$-barrel and flavodoxin-like folds it is possible to support homology under classic (*70*) sequence identity thresholds (figure 24). The alignment is structure-based, i.e. superimposable residues define the alignment and there is no further optimization to improve the sequence identity score. From the classical view of homology inference based on protein sequence identity: chance pairwise sequence identity goes up to 12%; if it is assumed that all residues can occur at equal frequency in proteins then it drops to 6% (*10*). Also, it has been argued that "above a cut-off roughly corresponding to 30% sequence identity, 90% of the pairs are homologous" and that "from 100–35% sequence identity, any residue exchange resulting in a stable structure maintains structure" (*78*).

Observing the identities depicted in table 15, the structure of NTM0182 clearly is a sequence identity bridge between both folds: in the global alignment it is more than 20% identical to both folds, while the alignment between the (β/α)$_8$-barrel and the flavodoxin-like folds is below 10% sequence identity. The intermediate features of NTM0182 are even more evident when looking at the scores of the αβαβ element aligned among the three structures: it shares 37% sequence identity, over 40 residues aligned, with both folds; while the same fragment between the (β/α)$_8$-barrel and the flavodoxin-like folds is only 18% identical.

37% sequence identity over a 40 residues alignment is clearly above the classical threshold (*70*) for supporting homology between pairs of proteins. Moreover, the possibility of superimposing this fragment with high scores in multiple homologous (β/α)$_8$-barrel superfamilies, and multiple times in the same structures already proven to have 2 and 3 fold symmetry are evidences of its intermediate properties. We suspect functional reasons for the conservation of the αβαβ element. It is likely that the three folds are binding a ligand with a phosphate moiety in the same region. The homology is well supported locally; the rest of the sequence between all the folds is less identical (aligned only by profile-profile comparisons). Perhaps the regions outside of the conserved element diverged by random drift or accommodated to the new structural context

As a summary of this section of our work we conclude that the HMM-HMM comparisons provided initial evidences for homology between different folds. The initial observations allowed us to focus on a reduced area of the sequence space to find a protein that displays intermediate features between the $(\beta/\alpha)_8$-barrel and the flavodoxin-like fold. The structure of NTM0182 revealed a very conserved structural motif that is a valuable evidence of common ancestry between the folds. The homology is now supported in classical terms (high sequence identity) in addition to the HHsearch probability scores. Since we recognized the potential of the HHsearch tool, we therefore employed it to compare all the folds classified in the SCOP α/β class.

4.4 Extending the profile-profile comparisons to all SCOP αβ folds.

As previously stated by Levitt "partitioning the protein structure universe into discrete folds does not exclude the similarities between protein domains that occur in different folds" (*4*). Protein classifications may be masking meaningful evolutionary relationships. Previously we already recognized the similarities between sub-structures from different folds; based on this observation we outlined a methodology that could be useful in diverse applications (*79*).

Subsequently, we explored the structure space using a sequence-based tool. We found many well-supported relationships among the most basal folds of the phylogenomic tree of architectures. We envisioned that our work could be used as a source of combinable fragments related by evolution that can be put together in order to expand the functionalities of the fold space. Being aware of the complication that carries the generation of chimeric proteins we decided to first build and test a novel chimera to see whether the sequence based fragments can be indeed combined in novel structural contexts.

4.5 Applied lessons from protein evolution: chimera building by combination of homologous fold-fragments.

We were able to build a well-folded chimera with native-like properties by combining fragments from the proteins LBP and CheY. The periplasmic binding protein I like fold resembles a duplication of the flavodoxin-like fold. In fact, the structural superposition of CheY on one of the lobes of LBP shows very good scores. It is possible to superimpose the same flavodoxin-like fold on the other lobe; however, the similarities are not as good. It is likely that the periplasmic binding protein evolved by a duplication of the flavodoxin-like fold, since the latter is more close to the base of the phylogenetic tree of architectures. We assume that combining related folds provides a good starting point for protein design.

The biophysical test performed on the chimera demonstrated that it is well folded. However, we did not manage to obtain an atomic model of the construct. Our redesign efforts were focused on the loops close to the hinge region. Moreover, we tried to use the Rosetta algorithm to evaluate how different designs may produce folded proteins. It is not clear whether the relaxation rounds performed with Rosetta really provided useful information. We discovered that the starting model we used as input to Rosetta has a strong influence on the final outcome. Even generating distinct homology models just by small variations in the alignment of template and target, already showed random variations in the calculated energies from Rosetta. We only observed some indications of problems with very bad energies when we used the closed conformation of LBP to build and minimize a homology model. In principle, Rosetta would evaluate all the models generated, with an open LBP, in the same way. No big differences were observed between the varieties of models used as input for the computation.

An atomic resolution model of the chimera would be very valuable since we could infer how the fragments from the different folds are fitting together. Especially we could observe if the fragments will stay in the same conformation or if they suffer big rearrangements to adapt to the novel context.

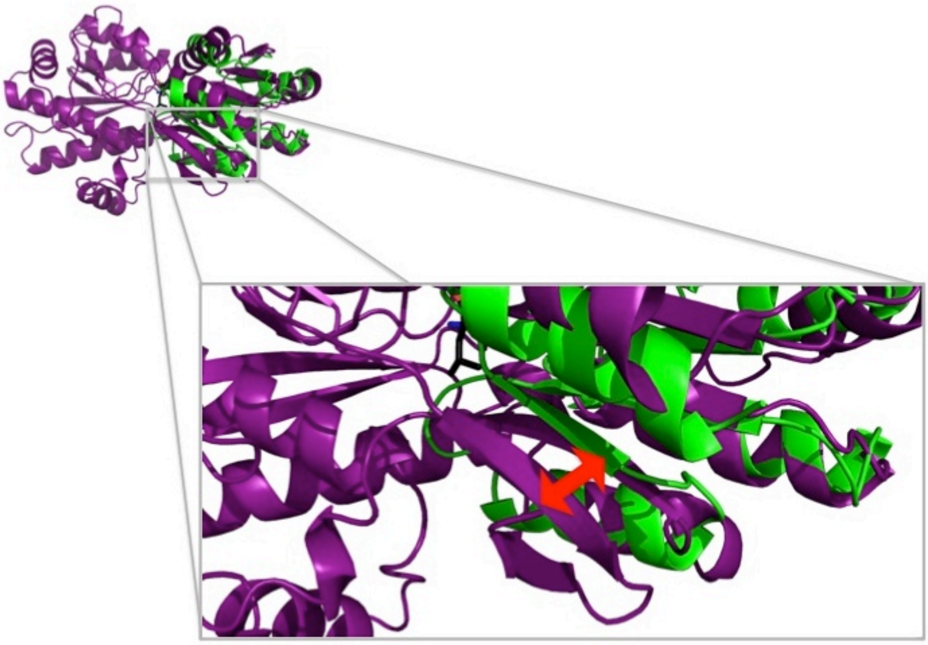


Fig. 43: Structural superposition of LBP (violet) and CheY (green). The area where optimization can improve the packing of chimera LBP-CheY is highlighted in gray. The red arrow shows where the β-strands from the different folds will have to establish interactions to yield a chimeric β-sheet.

Finally, we observed another area that may be problematic for the proper packing of the chimera LBP-CheY. β13 from LBP must establish interactions with β5 of CheY in order to build the chimeric β-sheet (figure 43). In the future those areas could be improved by protein design or directed evolution to design a better-packed chimera that will yield better crystals.

5. Final considerations

It is very likely that during the course of evolution modern protein domains were assembled by combination of smaller fragments. By comparing the sequences and structures of divergent pairs, we discover how nature designed the protein structure universe. The organization of protein structures into folds gives meaning to all the embellished shapes of the protein domains. Yet, the fold concept being arbitrary, it poses certain drawbacks. The protein classifications may obscure similarities in fold space and the comparison of sub-structures make the development of a metric of fold similarity almost unmanageable.

With our work we demonstrate that protein folds are homologous although they are inconsistently classified. Yet, the concept helps to infer evolutionary relationships. And since multiple times protein evolution studies have shown that modern protein domains evolved by combination of smaller fragments, we could employ the fold concept in conjunction with novel homology detection tools to discover previously unseen relationships.

Two general strategies can be outlined from our research. We first compare profiles that represent the structure space to find meaningful relationships. These comparisons allow us to focus on regions of the sequence space that still lack representative structures. Determination of intermediate structures at the boundaries of different fold classes will be very informative for our understanding of how sequence defines structure. Moreover, by this we can also contribute to map the structure space and learn about the evolutionary roads that might have led to fold change during evolution.

An enhanced knowledge of sequence-structure relationships will improve our ability to predict structures and design novel proteins with extended functionalities. We will have to rethink the way protein classifications are constructed to reflect the emergent view of a highly related protein fold space.

6. Summary of related publications and contributions

6.1 Evolutionary relationships of ancient superfolds.

In the past, common ancestry relationships between very divergent proteins were challenging to detect. Therefore, proteins that adopt different folds were classically considered to be non-homologous. Using state-of-the-art tools for homology detection I was able to find evidence for homology between proteins of two ancestral superfolds, the $(\beta\alpha)_8$-barrel (or TIM-barrel) and the flavodoxin-like fold. Moreover, I identified a family of proteins that showed intermediate features between both folds. I determined the first crystal structure of one of its members. This atomic model provided novel insights into the evolutionary roads followed by two of the earliest folds. Our combined insights provide support for an emergent vision where protein superfolds share common ancestry.

Evolutionary relationship of two ancient protein superfolds
Farías-Rico JA & Höcker B.
Submitted to Nature, under review.

6.2 Design of chimeric proteins by combination of subdomain-sized fragments

As a natural extension of the superfold evolution project, we searched the protein structure space to find additional homologous folds. We discovered evidences of homology between the periplasmic binding protein-like I and the flavodoxin-like fold. To gain insight into how evolution might have occurred, we combined fragments that originated from these two folds. Having these experiments as background and guidance, but without including the sequence-based novelty approach, we wrote the following methods paper.

Methods Enzymol. 2013;523:389-405
Design of chimeric proteins by combination of subdomain-sized fragments.
Farías-Rico JA, Höcker B.

6.3 Change in protein-ligand specificity through binding pocket grafting.

The periplasmic binding proteins are promising scaffolds for biosensor engineering. In our lab, we are interested in understanding the mechanisms underlying protein-binding specificity. PotF and PotD, periplasmic binding proteins, show distinctive specificities for putrescine and spermidine. By mutating 7 residues in PotF to the ones present in the binding pocket of PotD, the ligand specificity of the protein was swapped. Our work shows how it is possible to successfully change the protein-ligand specificity through transplanting the binding pocket from PotD onto PotF. Moreover, we show how the specificity is encoded in the first shell residues of the PotF binding pocket. In this project, I was involved in the determination of the PotF mutant crystal structure.

J Struct Biol. 2013 Jun 17. pii: S1047-8477(13)00157-3.
Change in protein-ligand specificity through binding pocket grafting.
Scheib U, Shanmugaratnam S, **Farías-Rico JA**, Höcker B.

6.4 Detailed description of my contributions to this dissertation.

Under the guidance and support from my advisor Birte Höcker, I performed all the work described in this doctoral thesis. I supervised Saacnicteh Toledo Patiño, an undergraduate student working in our laboratory, who contributed to this work in the following ways: a) cloning, expression and purification of some of the intermediate candidates between different folds, b) cloning, expression and purification of some chimeric constructs, c) biophysical tests of some of the proteins experimentally tested. Saacnicteh Toledo Patiño and Matthias Schwer developed a computational algorithm to parse the sequence data that I generated by the HMM-HMM comparisons. This algorithm made the detection of the sequence-based connection among the three folds addressed in this work possible. Steffen Schmidt contributed with valuable bioinformatic support and plotted figures 8 and 9.

7. References

1. F. J. Ayala, "Nothing in biology makes sense except in the light of evolution": Theodosius Dobzhansky: 1900-1975. *The Journal of heredity* **68**, 3 (Jan-Feb, 1977).
2. H. S. Subramanya, A. J. Doherty, S. R. Ashford, D. B. Wigley, Crystal structure of an ATP-dependent DNA ligase from bacteriophage T7. *Cell* **85**, 607 (May 17, 1996).
3. K. Takeda *et al.*, Crystal structure of the M intermediate of bacteriorhodopsin: allosteric structural changes mediated by sliding movement of a transmembrane helix. *Journal of molecular biology* **341**, 1023 (Aug 20, 2004).
4. R. Kolodny, L. Pereyaslavets, A. O. Samson, M. Levitt, On the universe of protein folds. *Annual review of biophysics* **42**, 559 (2013).
5. M. Magrane, U. Consortium, UniProt Knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation* **2011**, bar009 (2011).
6. M. Punta *et al.*, The Pfam protein families database. *Nucleic acids research* **40**, D290 (Jan, 2012).
7. A. Bateman, P. Coggill, R. D. Finn, DUFs: families in search of function. *Acta crystallographica. Section F, Structural biology and crystallization communications* **66**, 1148 (Oct 1, 2010).
8. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* **247**, 536 (Apr 7, 1995).
9. F. M. Pearl *et al.*, The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic acids research* **31**, 452 (Jan 1, 2003).
10. R. F. Doolittle, Similar amino acid sequences: chance or common ancestry? *Science* **214**, 149 (Oct 9, 1981).
11. E. V. Koonin, Y. I. Wolf, G. P. Karev, The structure of the protein universe and genome evolution. *Nature* **420**, 218 (Nov 14, 2002).
12. A. G. Murzin, How far divergent evolution goes in proteins. *Current opinion in structural biology* **8**, 380 (Jun, 1998).
13. S. Rowsell *et al.*, Crystal structure of carboxypeptidase G2, a bacterial enzyme with applications in cancer therapy. *Structure* **5**, 337 (Mar 15, 1997).
14. A. N. Lupas, C. P. Ponting, R. B. Russell, On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of structural biology* **134**, 191 (May-Jun, 2001).
15. N. V. Grishin, Fold change in evolution of protein structures. *Journal of structural biology* **134**, 167 (May-Jun, 2001).
16. G. Caetano-Anolles, D. Caetano-Anolles, An evolutionarily structured universe of protein architecture. *Genome research* **13**, 1563 (Jul, 2003).
17. C. A. Orengo, D. T. Jones, J. M. Thornton, Protein superfamilies and domain superfolds. *Nature* **372**, 631 (Dec 15, 1994).
18. R. D. Hills, Jr. *et al.*, Topological frustration in beta alpha-repeat proteins: sequence diversity modulates the conserved folding mechanisms of alpha/beta/alpha sandwich proteins. *Journal of molecular biology* **398**, 332 (Apr 30, 2010).
19. J. Yuan, R. W. Branch, B. G. Hosu, H. C. Berg, Adaptation at the output of the chemotaxis signalling pathway. *Nature* **484**, 233 (2012).
20. C. L. Drennan, S. Huang, J. T. Drummond, R. G. Matthews, M. L. Lidwig, How a protein binds B12: A 3.0 A X-ray structure of B12-binding domains of methionine synthase. *Science* **266**, 1669 (Dec 9, 1994).
21. G. Caetano-Anolles, H. S. Kim, J. E. Mittenthal, The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 9358 (May 29, 2007).

22. E. D. Nelson, N. V. Grishin, Alternate pathways for folding in the flavodoxin fold family revealed by a nucleation-growth model. *Journal of molecular biology* **358**, 646 (May 5, 2006).

23. Y. J. Bollen, C. P. van Mierlo, Protein topology affects the appearance of intermediates during the folding of proteins with a flavodoxin-like fold. *Biophysical chemistry* **114**, 181 (Apr 22, 2005).

24. B.-G. Ma *et al.*, Characters of very ancient proteins. *Biochemical and Biophysical Research Communications* **366**, 607 (2008).

25. H. Szurmant, G. W. Ordal, Diversity in chemotaxis mechanisms among the bacteria and archaea. *Microbiology and molecular biology reviews : MMBR* **68**, 301 (Jun, 2004).

26. R. K. Wierenga, The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS letters* **492**, 193 (Mar 16, 2001).

27. D. Lang, R. Thoma, M. Henn-Sax, R. Sterner, M. Wilmanns, Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* **289**, 1546 (Sep 1, 2000).

28. M. C. Vega, E. Lorentzen, A. Linden, M. Wilmanns, Evolutionary markers in the (beta/alpha)8-barrel fold. *Current opinion in chemical biology* **7**, 694 (Dec, 2003).

29. L. Jiang *et al.*, De novo computational design of retro-aldol enzymes. *Science* **319**, 1387 (Mar 7, 2008).

30. R. R. Copley, P. Bork, Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *Journal of molecular biology* **303**, 627 (Nov 3, 2000).

31. N. Nagano, C. A. Orengo, J. M. Thornton, One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of molecular biology* **321**, 741 (Aug 30, 2002).

32. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389 (Sep 1, 1997).

33. J. Soding, Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951 (Apr 1, 2005).

34. V. Alva, M. Remmert, A. Biegert, A. N. Lupas, J. Soding, A galaxy of folds. *Protein science : a publication of the Protein Society* **19**, 124 (Jan, 2010).

35. T. Frickey, A. Lupas, CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702 (Dec 12, 2004).

36. B. Hocker, S. Schmidt, R. Sterner, A common evolutionary origin of two elementary enzyme folds. *FEBS letters* **510**, 133 (Jan 16, 2002).

37. T. A. Bharat, S. Eisenbeis, K. Zeth, B. Hocker, A beta alpha-barrel built by the combination of fragments from different folds. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 9942 (Jul 22, 2008).

38. A. Sali, Comparative protein modeling by satisfaction of spatial restraints. *Molecular medicine today* **1**, 270 (Sep, 1995).

39. C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, Protein structure prediction using Rosetta. *Methods in enzymology* **383**, 66 (2004).

40. S. Eisenbeis *et al.*, Potential of fragment recombination for rational design of proteins. *Journal of the American Chemical Society* **134**, 4019 (Mar 7, 2012).

41. M. A. Dwyer, H. W. Hellinga, Periplasmic binding proteins: a versatile superfamily for protein engineering. *Current opinion in structural biology* **14**, 495 (Aug, 2004).

42. R. Tam, M. H. Saier, Jr., Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiological reviews* **57**, 320 (Jun, 1993).

43. F. A. Quiocho, P. S. Ledvina, Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. *Molecular microbiology* **20**, 17 (Apr, 1996).

44. U. Scheib, S. Shanmugaratnam, J. A. Farias-Rico, B. Hocker, Change in protein-ligand specificity through binding pocket grafting. *Journal of structural biology*, (Jun 17, 2013).

45. D. Wu *et al.*, Structural basis of substrate binding specificity revealed by the crystal structures of polyamine receptors SpuD and SpuE from Pseudomonas aeruginosa. *Journal of molecular biology* **416**, 697 (Mar 9, 2012).

46. B. Hocker, J. Claren, R. Sterner, Mimicking enzyme evolution by generating new (betaalpha)8-barrels from (betaalpha)4-half-barrels. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16448 (Nov 23, 2004).

47. http://www.lifetechnologies.com/de/de/home/life-science/cloning/gene-synthesis.html.

48. G. Wu *et al.*, Simplified gene synthesis: a one-step approach to PCR-based gene construction. *Journal of biotechnology* **124**, 496 (Jul 25, 2006).

49. J. A. Bornhorst, J. J. Falke, Purification of proteins using polyhistidine affinity tags. *Methods in enzymology* **326**, 245 (2000).

50. A. C. Grodzki, E. Berenstein, Antibody purification: ion-exchange chromatography. *Methods in molecular biology* **588**, 27 (2010).

51. L. Hagel, Gel-filtration chromatography. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 10**, Unit 10 9 (May, 2001).

52. U. K. Laemmli, Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680 (Aug 15, 1970).

53. E. Khazina, O. Weichenrieder, Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 731 (Jan 20, 2009).

54. R. L. Krauth-Siegel *et al.*, Crystallization and preliminary crystallographic analysis of trypanothione reductase from Trypanosoma cruzi, the causative agent of Chagas' disease. *FEBS letters* **317**, 105 (Feb 8, 1993).

55. G. M. Sheldrick, A short history of SHELX. *Acta crystallographica. Section A, Foundations of crystallography* **64**, 112 (Jan, 2008).

56. L. M. Rice, T. N. Earnest, A. T. Brunger, Single-wavelength anomalous diffraction phasing revisited. *Acta crystallographica. Section D, Biological crystallography* **56**, 1413 (Nov, 2000).

57. G. Bricogne, C. Vonrhein, C. Flensburg, M. Schiltz, W. Paciorek, Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta crystallographica. Section D, Biological crystallography* **59**, 2023 (Nov, 2003).

58. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* **66**, 213 (Feb, 2010).

59. K. Cowtan, The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta crystallographica. Section D, Biological crystallography* **62**, 1002 (Sep, 2006).

60. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography* **60**, 2126 (Dec, 2004).

61. E. Krissinel, K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta crystallographica. Section D, Biological crystallography* **60**, 2256 (Dec, 2004).

62. J. Soding, A. Biegert, A. N. Lupas, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**, W244 (Jul 1, 2005).

63. https://rosettacommons.org/home.

64. http://wiki.lesswrong.com/wiki/Bayesian_probability.

65. P. A. Alexander, Y. He, Y. Chen, J. Orban, P. N. Bryan, The design and characterization of two proteins with 88% sequence identity but different structure

and function. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11963 (Jul 17, 2007).

66. S. Dalal, S. Balasubramanian, L. Regan, Protein alchemy: changing beta-sheet into alpha-helix. *Nature structural biology* **4**, 548 (Jul, 1997).

67. C. G. Roessler *et al.*, Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2343 (Feb 19, 2008).

68. A. Biegert, C. Mayer, M. Remmert, J. Soding, A. N. Lupas, The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic acids research* **34**, W335 (Jul 1, 2006).

69. V. B. Chen *et al.*, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica. Section D, Biological crystallography* **66**, 12 (Jan, 2010).

70. C. Sander, R. Schneider, The HSSP data base of protein structure-sequence alignments. *Nucleic acids research* **21**, 3105 (Jul 1, 1993).

71. P. Alifano *et al.*, Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiological reviews* **60**, 44 (Mar, 1996).

72. J. Soding, M. Remmert, A. Biegert, HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic acids research* **34**, W137 (Jul 1, 2006).

73. C. Chothia, Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543 (Jun 18, 1992).

74. R. N. McLaughlin, Jr., F. J. Poelwijk, A. Raman, W. S. Gosal, R. Ranganathan, The spatial architecture of protein function and adaptation. *Nature* **491**, 138 (Nov 1, 2012).

75. J. Huang, A. Koide, K. Makabe, S. Koide, Design of protein function leaps by directed domain interface evolution. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6578 (May 6, 2008).

76. N. Koga *et al.*, Principles for designing ideal protein structures. *Nature* **491**, 222 (Nov 8, 2012).

77. M. Remmert, A. Biegert, A. Hauser, J. Soding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **9**, 173 (Feb, 2012).

78. B. Rost, Twilight zone of protein sequence alignments. *Protein engineering* **12**, 85 (Feb, 1999).

79. J. A. Rico, B. Hocker, Design of chimeric proteins by combination of subdomain-sized fragments. *Methods in enzymology* **523**, 389 (2013).

8. Appendix

8.1 Homology modeling pairwise alignment

Pir format for Modeller input

```
>P1;LBP-CheY-V5/1-340
sequence:LBP-CheY-V5:   1: : 340: :: : 0.00: 0.00
MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIK
YVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQRVLI
VDDAAFM----RMMLKDIITKAGYEVAG------EATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEI
MKIDPNAKIIVGPEGVGNQAMVIEAIKAGAKDLVTMPKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAV
QSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKGFDFGVFQWHADGSSTKAK*
>P1;1usk/1-346
structureX:1usk: :A: :A:c.93.1.1 (A) Leucine-binding protein [Thermotoga
maritima]:1.53:0.17
-DDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIK
YVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQRIAI
IHDKQQYGEGLARSVQDGLKAANANVVFFDGITAGEKDFSALIARLKKENIDFVYY--GGYYPEMGQMLRQA
RSVGLKTQ-FMGPEGVGNASLSNIAGDAAEGMLVTMPKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAV
QSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKGFDFGVFQWHADGSSTKAK*
>P1;1u0s/1-110
structureX:1u0s: :Y: :Y:c.23.1.1 (Y) CheY protein [Thermotoga
maritima]:1.90:0.23
------------------------------------------------------------------------
------------------------------------------------------------------GKRVLI
VDDAAFM----RMMLKDIITKAGYEVAG------EATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEI
MKIDPNAK-IIVCSAMGQQAMVIEAIKAGAKDFIVKPFQPSRV-----------------------------
--------------------------------------------------------------*
```

8.2 Topology cartoons of chimeras LBP-CheY 2 to 5

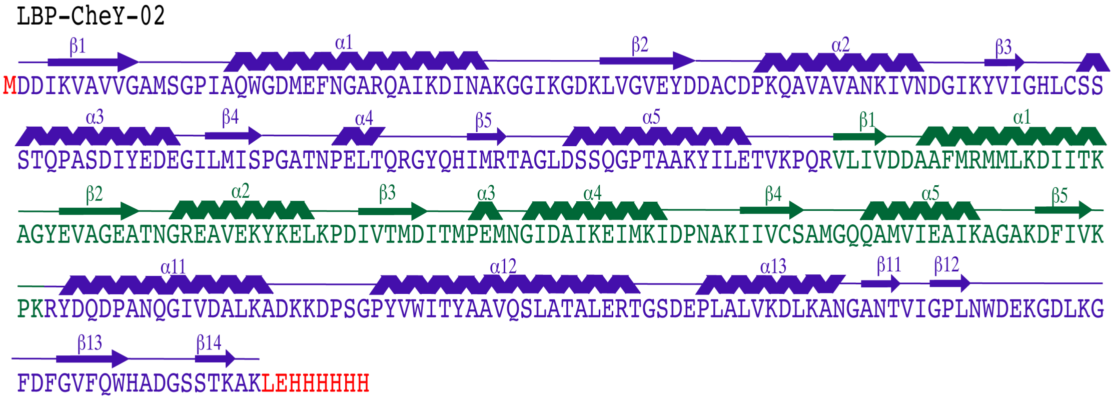

LBP-CheY-02
β1
α1
β2
α2
β3
MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSS
α3
β4
α4
β5
α5
β1
α1
STQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQRVLIVDDAAFMRMMLKDIITK
β2
α2
β3
α3
α4
β4
α5
β5
AGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVK
α11
α12
α13
β11
β12
PKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKG
β13
β14
FDFGVFQWHADGSSTKAKLEHHHHHH

Fig. 44: LBP-CheY-02 topology cartoon

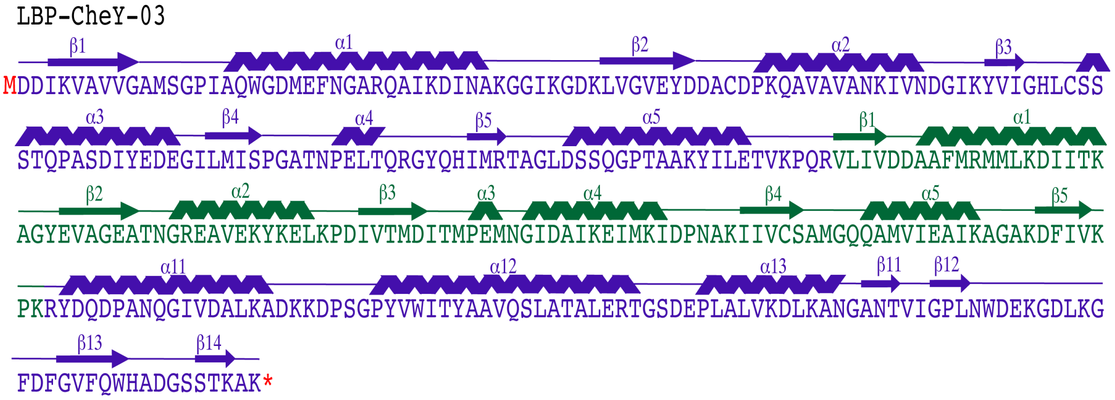

LBP-CheY-03
β1
α1
β2
α2
β3
MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSS
α3
β4
α4
β5
α5
β1
α1
STQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQRVLIVDDAAFMRMMLKDIITK
β2
α2
β3
α3
α4
β4
α5
β5
AGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVK
α11
α12
α13
β11
β12
PKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKG
β13
β14
FDFGVFQWHADGSSTKAK*

Fig. 45: LBP-CheY-03 topology cartoon

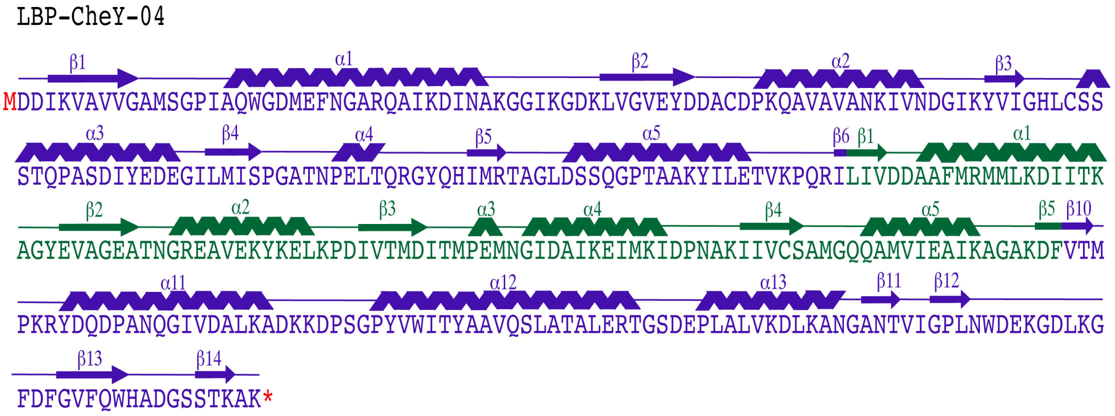

LBP-CheY-04
β1
α1
β2
α2
β3
MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSS
α3
β4
α4
β5
α5
β6
β1
α1
STQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQRILIVDDAAFMRMMLKDIITK
β2
α2
β3
α3
α4
β4
α5
β5
β10
AGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFVTM
α11
α12
α13
β11
β12
PKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKG
β13
β14
FDFGVFQWHADGSSTKAK*

Fig. 46: LBP-CheY-04 topology cartoon

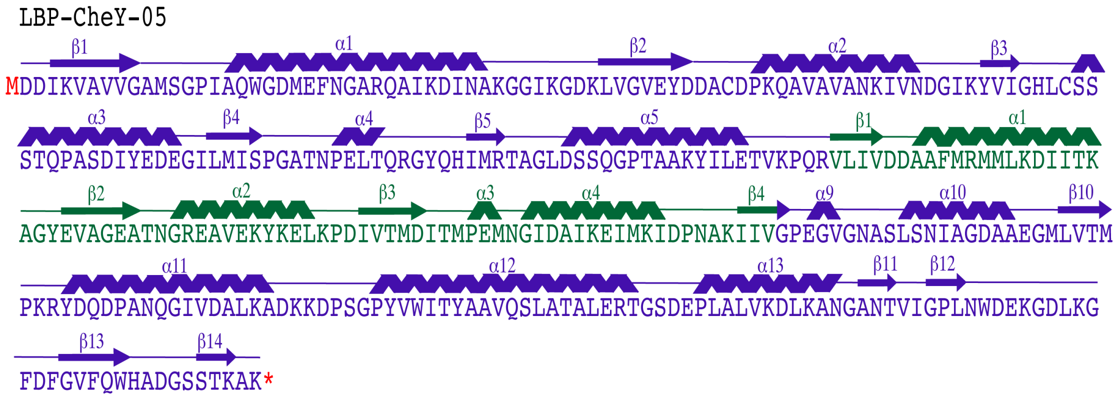

LBP-CheY-05
β1
α1
β2
α2
β3
MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSS
α3
β4
α4
β5
α5
β1
α1
STQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQRVLIVDDAAFMRMMLKDIITK
β2
α2
β3
α3
α4
β4
α9
α10
β10
AGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVGPEGVGNASLSNIAGDAAEGMLVTM
α11
α12
α13
β11
β12
PKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKG
β13
β14
FDFGVFQWHADGSSTKAK*

Fig. 47: LBP-CheY-05 topology cartoon

8.3 Sequences, primers and sources of experimentally tested intermediate sequences

Table 22: Sequences of intermediate targets experimentally characterized.

gi|15642956|ref|NP_227997.1| hypothetical protein TM0182 [*Thermotoga maritima* MSB8]	
Full sequence	MYILFREMKNNWYSLAALLSTIYSRHLDVEARPVKFEEIKKFPPEKTIVAYSFMSFDLDTVREEVKTLKERGYTLIAGGPHVTADPEGCLRMGFDHVFTGDGEENILKFLMGERKKIFDGISKRVNLNHYPPFLPSKGIYMPIEITRGCPFSCAYCQTPIIAGRRVRHRDVDVVVHYAKLGVKHGRKLARFIAPNSFGYGSKNGVTPNVEKIEELLYGLKKVGIEEIYFGTFPSEVRPESVTDEVLKVVKKYVNNRSIVIGAQSGSDRILKIIKRGHTVEQVEEAIEKISLHGFIPHVDFIFGFPFETEEDVEKTFSFIVKIVERYGAKIHAHTFMPLPGTELFNAGPGRLTEVHYKFLGRLASKGILDGYWMKQEMLARKVYEIASGGSTDVTSDR
TM01	MYILFREMKNNWYSLAALLSTIYSRHLDVEARPVKFEEIKKFPPEKTIVAYSFMSFDLDTVREEVKTLKERGYTLIAGGPHVTADPEGCLRMGFDHVFTGDGEENILKFLMGERKKIFDGLEHHHHHH
Primers	>156429561_THERMO_FW CGCATATGTATATTCTCTTCAGAGAGATGAAG >15642956120_THERMO_RV CCTCGAGACCATCGAAGATCTTTTTCTCTCC
Gene source	Genomic DNA
TM02	MYILFREMKNNWYSLAALLSTIYSRHLDVEARPVKFEEIKKFPPEKTIVAYSFMSFDLDTVREEVKTLKERGYTLIAGGPHVTADPEGCLRMGFDHVFTGDGEENILKFLMGLEHHHHHH
Primers	>156429561_THERMO_FW CGCATATGTATATTCTCTTCAGAGAGATGAAG >15642956120_THERMO_RV_MOD CCTCGAGCCCCATCAAGAACTTCAAAATATTCTC
Gene source	Plasmid pET-21-TM01 AmpR

gi|254458087|ref|ZP_05071514.1| methylaspartate mutase, S subunit [*Campylobacterales bacterium* GD 1]		
Full sequence		MIELSLNARGFEVFNLGVNTYLEEFFDAVVETGADILLISSLNGEAEGWSREIKLLKSKYKNLDNLVMMIGGNLVVGSADAETIIPKYKNYGFDLVFHQVDLNTGLDTLEEFLKERNK
CM01		MIELSLNARGFEVFNLGVNTYLEEFFDAVVETGADILLISSLNGEAEGWSREIKLLKSKYKNLDNLVMMIGGNLVVGSADAETIIPKYKNYGFDLVFHQVDLNTGLDTLEEFLKERNKELEHHHHHH
	Sub-cloning	*NdeI, XhoI*
	Gene source	Purchased plasmid 0946287_254458087_M_mutase_pMA AmpR
CM02[1]		MKVVTGVVGNDIHVVANRLIELSLNARGFEVFNLGVNTYLEEFFDAVVETGADILLISSLNGEAEGWSREIKLLKSKYKNLDNLVMMIGGNLVVGSADAETIIPKYKNYGFDLVFHQVDLNTGLDTLEEFLKELEHHHHHH
	Primers	>fw-camp-Nter CGCATATGAAAGTAGTAACAGGCGTAGTCGGAAATGACATTCATGTTGTAGCAAATAGATTGATCGAGCTGTCGCTGAAT >rv-camp-Cter CCTCGAGCTCTTTCAGAAACTCCTCCAGTGT
	Gene source	Plasmid pET-21-CM01 AmpR

[1] ORF corrected in NCBI database, gi|298283453|gb|EDZ62765.2| *methylaspartate mutase*, S subunit [Campylobacterales bacterium GD 1] Version 2 (cloned) CM02

gi|15897403|ref|NP_342008.1| hypothetical protein SSO0477 [*Sulfolobus solfataricus* P2]		
Full sequence		MNTFYSRKVKYNIMAWEIILTADKGSFTDYGGSSVLGYVACMPSRLIPKFFMDRFFTPDVPVDSEGRAIVAPYALRKVESTLVHAGFDSVVVIPPHRLEKAINQKTKVVGLTVHDPFGLNPVSFKLSMIFGGGPTWTAKYFEEFGEKISKLKSKYNFKVIVGGPGSWELTKENKDWADVIFIGEAEADLPRVVKSIIDGQEVPKVVYGKNPKVNEIPPIINPARLGEVQITRGCPRGCQFCPITPETFRTIPLDVVKKEVEVNMRAGVKRVEFITDDVLLYGSQKLRVNHEAITKLFTETMNMGVDGIWFPHISAPAVRSSPQTVKAMSEIARYDEDRAAAPVVGLESGSEKILSKYMRAKPFPWTPREWKDVILDATAIMNDNYIYPCYTMTIGYPEETNEDVDQSIDLVQSIIDHKLKAWIFPLPVIPMGVSYIRNNPFPVLEKMPTRYWDVLYISWKYDLQITREMIPILTGGIKNKFAQRTVQYMIDKIFYSIEWVFKQLKETQGKYAYTFASINLNNTTGVIKAIYWLFRLAFKPL
SS01		MAWEIILTADKGSFTDYGGSSVLGYVACMPSRLIPKFFMDRFFTPDVPVDSEGRAIVAPYALRKVESTLVHAGFDSVVVIPPHRLEKAINQKTKVVGLTVHDPFGLNPVSFKLSMIFGGGPTWTAKYFEEFGEKISKLKSKYNFKVIVGGPGSWELTKENKDWADVIFIGEAEALEHHHHHH
	Primers	>15897403_ss_N-dom14FW TGCATATGATGGCATGGGAGAGATCATATTAACC >15897403_ss_N-dom187RV TCTCGAGTGCTTCAGCTTCTCCTATGAATATAAC
	Gene source	Genomic DNA
SS02		MNTFYSRKVKYNIMMAWEIILTADKGSFTDYGGSSVLGYVACMPSRLIPKFFMDRFFTPDVPVDSEGRAIVAPYALRKVESTLVHAGFDSVVVIPPHRLEKAINQKTKVVGLTVHDPFGLNPVSFKLSMIFGGGPTWTAKYFEEFGEKISKLKSKYNFKVIVGGPGSWELTKENKDWADVIFIGEAEALEHHHHHH
	Primers	>15897403_ss_N-dom14FW_mod TGCATATGGCATGGGAGATCATATTAACCGC >15897403_ss_N-dom187RV TCTCGAGTGCTTCAGCTTCTCCTATGAATATAAC
	Gene source	Genomic DNA
SS03		MNTFYSRKVKYNIMAWEIILTADKGSFTDYGGSSVLGYVACMPSRLIPKFFMDRFFTPDVPVDSEGRAIVAPYALRKVESTLVHAGFDSVVVIPPHRLEKAINQKTKVVGLTVHDPFGLNPVSFKLSMIFGGGPTWTAKYFEEFGEKISKLKSKYNFKVIVGGPGSWELTKENKDWADVIFIGEAEADLPRVVKSIILEHHHHHH
	Primers	>fw-15897403-v3.0 TGCATATGAACACCTTTTATAGCCGT >rv-15897403-v3.0 TCTCGAGAATTATGGATTTTACAACTCGTGGTAGATCTGCTTCAGCTTCTCCTATGAATATAACATC
	Gene source	pET-21-SS02

gi|77918738|ref|YP_356553.1| putative Fe-S oxidoreductase [*Pelobacter carbinolicus* DSM 2380]		
Full sequence		MNYLFVVPRDNCFGFLTIPPGVVYVATSLKETGRNVYGIHLNYESDTKESLKKKIIDNNIDVLCIGGLSDQYNEIKRTIDLSKQIKPDLIIVVGGGLITAQPTLIMENIGADYAIVGQGEITICELAEALEGKKPIRDVAGIVYFENQALVCNENRPEIRELDTVTNPDYDIFPYTQVPNDPININGDFKRTVNITASRSCPYNCTFCYHPSGTTYRQRSIENIFREIDFLLSKYDIEHLLIIDELFAIDENRVSEFCEAIAKYDVTFSVQLRVDGIDENLLLKLKNAGCTSISYGLESADNSILKSMKKGTDISQIEKALSLTRKIGFFIQGYFIFGDIEETMGTVNTTIRWWMKHLEYGINLAMIRIFPGSYLYQHAIEQSIITDQMQYIENGCPLINISKLTDQEFAGLIKKSPILIRNFSV
PC01		MNYLFVVPRDNCFGFLTIPPGVVYVATSLKETGRNVYGIHLNYESDTKESLKKKIIDNNIDVLCIGGLSDQYNEIKRTIDLSKQIKPDLIIVVGGGLITAQPTLIMENIGADYAIVGQGEITICELAEALEGKLEHHHHHH
	Sub-cloning	*NdeI, XhoI*
	Gene source	Plasmid 1 015643.gb FeSoxPelobacter pMA-T AmpR
PC02		MIPPGVVYVATSLKETGRNVYGIHLNYESDTKESLKKKIIDNNIDVLCIGGLSDQYNEIKRTIDLSKQIKPDLIIVVGGGLITAQPTLIMENIGADYAIVGQGEITICELAEALEGKLEHHHHHH
	Primers	>fw-peloprimer1 CGCATATGGGCCGTAACGTCTATGGAATTCACCTGAACTATGAG >rv-peloprimer GCTCGAGTTTGCCTTCCAGAGCTTCAGCCAGTTCACAGAT
	Gene source	pET-21-PC01
PC03		MGRNVYGIHLNYESDTKESLKKKIIDNNIDVLCIGGLSDQYNEIKRTIDLSKQIKPDLIIVVGGGLITAQPTLIMENIGADYAIVGQGEITICELAEALEGKLEHHHHHH
	Primers	>fw-pelo-primer2 CGCATATGATTCCGCCTGGAGTAGTGTATGTGGCCACCTCA >rv-peloprimer GCTCGAGTTTGCCTTCCAGAGCTTCAGCCAGTTCACAGAT
	Gene source	pET-21-PC01
PC04		MNYLFVVPRDNCFGFLTIPPGVVYVATSLKETGRNVYGIHLNYESDTKESLKKKIIDNNIDVLCIGGLSDQYNEIKRTIDLSKQIKPDLIIVVGGGLITAQPTLIMENIGADYAIVGQGEITICELAEALEGKKPIRDVAGIVYFENQALVLEHHHHHH
	Primers	>FW-PELO-v1.0 CGCATATGAACTATCTGTTTGTGGTG >Rv-PELO-v1.0 CCTCGAGGACAAGTGCTTGGTTTTCAAAATACACGATCCCGGCAACATCGCGGATCGGTTTTTTGCCTTCCAGAGCTTCAGCCAG
	Gene source	pET-21-PC01
PC05		MIPPGVVYVATSLKETGRNVYGIHLNYESDTKESLKKKIIDNNIDVLCIGGLSDQYNEIKRTIDLSKQIKPDLIIVVGGGLITAQPTLIMENIGADYAIVGQGEITICELAEALEGKKPIRDVAGIVYFENQALVLEHHHHHH
	Primers	>fw-pelo-primer2 CGCATATGATTCCGCCTGGAGTAGTGTATGTGGCCACCTCA >Rv-PELO-v1.0 CCTCGAGGACAAGTGCTTGGTTTTCAAAATACACGATCCCGGCAACATCGCGGATCGGTTTTTTGCCTTCCAGAGCTTCAGCCAG
	Gene source	pET-21-PC01

PC06	MGRNVYGIHLNYESDTKESLKKKIIDNNIDVLCIGGLSDQYNEIKRTIDLSKQIKPDLIIVVGGGLITAQPTLIMENIGADYAIVGQGEITICELAEALEGKKPIRDVAGIVYFENQALVLEHHHHHH
Primers	>fw-peloprimer1 CGCATATGGGCCGTAACGTCTATGGAATTCACCTGAACTATGAG >Rv-PELO-v1.0 CCTCGAGGACAAGTGCTTGGTTTTCAAAATACACGATCCCGGCAACATCGCGGATCGGTTTTTGCCTTCCAGAGCTTCAGCCAG
Gene source	pET-21-PC01

>gi|16272184|ref|NP_438393.1| inosine-5'-monophosphate dehydrogenase-like protein [*Haemophilus influenzae* Rd KW20]	
Full sequence	MTNIHYHKILILDFGSQYTQLIARRVREIGVYCELWAWDVTEQXIREFAPELY QGRAFKSYRGMGSLGAMAKGSSDRYFQSDNAADKLVPEGIEGRIPYKGYLK EIIHQQMGGLRSCMGLTGCATIDELRTKAEFVRISGAGIKESHVHDVAITKEA PNYRMG
Gene source	Gene assembly[1]
HI01	MTNIHYHKILILDFGSQYTQLIARRVREIGVYCELWAWDVTEQXIREFAPELY QGRAFKSYRGMGSLGAMAKGSSDRYFQSDNAADKLVPEGIEGRIPYKGYLK EIIHQQMGGLRSCMGLTGCATIDELRTKAEFVRISGAGIKESHVHDVAITKEA PNYRMGLEHHHHHH
Primers	Internal: >1_h.influenzae_sens ATGACCAACATTCATTATCATAAAATTCTGATTCTGGATTTTGGCAGCCA GTATACCCAG >2_h.influenzae_anti GTTCGCAATACACGCCAATTTCACGCACACGACGCGCAATCAGCTGGGTA TACTGGCTGC >3_h.influenzae_sens TTGGCGTGTATTGCGAACTGTGGGCGTGGGATGTGACCGAACAGCAGATT CGTGAATTTG >4_h.influenzae_anti ACGATAGCTTTTAAACGCACGGCCCTGATACAGTTCCGGCGCAAATTCAC GAATCTGCTG >5_h.influenzae_sens GTGCGTTTAAAAGCTATCGTGGCATGGGCAGCCTGGGCGCGATGGCGAA AGGCAGCAGCG >6_h.influenzae_anti CGGCACCAGTTTATCCGCCGCGTTATCGCTCTGAAAATAACGATCGCTGC TGCCTTTCGC >7_h.influenzae_sens CGGATAAACTGGTGCCGGAAGGCATTGAAGGCCGTATTCCGTATAAAGG CTATCTGAAAG >8_h.influenzae_anti ATGCAGCTACGCAGGCCGCCCATCTGCTGATGAATAATTTCTTTCAGATA GCCTTTATAC >9_h.influenzae_sens GGCCTGCGTAGCTGCATGGGCCTGACCGGCTGCGCGACCATTGATGAACT GCGTACCAAA >10_h.influenzae_anti CTTTCTTTAATGCCCGCGCCGCTAATACGCACAAATTCCGCTTTGGTACGC AGTTCATCA >11_h.influenzae_sens CGCGGGCATTAAAGAAAGCCATGTGCATGATGTGGCGATTACCAAAGAA GCGCCGAACTA
Primers	>12_h.influenzae_anti GCCCATACGATAGTTCGGCGCTTCTTTG External >h_in_fw CGCATATGACCAACATTCATTATCAT >h_in_rv CCTCGAGGCCCATACGATAGTTCG

[1] The gene was assembled using 12 internal primers and 2 external primers. The method was adapted from (*48*).

>gi|164688057|ref|ZP_02212085.1| hypothetical protein CLOBAR_01702 [*Clostridium bartlettii* DSM 16795]	
Full sequence	MSKRKQVTVPMEKIKEQDKYINEIKNENERYFHLTGKKKSYFIQTFGCQMNEHDSEKLGAMLNAMGYEPSLMADNADLIIYNTCAVRENAELKVYGNLGHLKLIKRRNPNLKIAVCGCMMQQPAIVKEIKAKYKHVDLVFGTHNLYKFPELLSESMSSDSILIDVWDVDGEVVEGLRSDRKFELKAFVNIMYGCNNFCTYCIVPYTRGRERSRRPEDIMNEIKELVANGTKEVTLLGQNVDSYGKTLEEEDRMTFAELLRAVNEIDGLERIRFMTSHPKDISDEVIYAMRDCDKVCEFLHLPVQCGSTKLLKKMNRHYSKEDYLRIVEKAKAEVPNIAFSTDIMVGFPGETEEDVEDTLDVIRQVRYDNAFTFIYSKRTGTPAAKMEDQIPEDVKHKRFNRVLELVNEISKENNTTHQDEVVEILVEGKSKTDDTKFTGRTRQNKLVNFSVKNPDADLIGKLVNVKITEAALSFSLNGEMVE
CB01	MKSYFIQTFGCQMNEHDSEKLGAMLNAMGYEPSLMADNADLIIYNTCAVRENAELKVYGNLGHLKLIKRRNPNLKIAVCGCMMQQPAIVKEIKAKYKHVDLVFGTHNLYKFPELLSESMSSLEHHHHHH
Primers	Internal >164688057-SS-1 ATGAAAAGCTATTTTATTCAGACCTTTGGCTGCCAGATGAACGAACATGATAGCGAAAAA >164688057-AS-2 GGCTCGGTTCATAGCCCATCGCGTTCAGCATCGCGCCCAGTTTTTCGCTATCATGTTCGT >164688057-SS-3 GGGCTATGAACCGAGCCTGATGGCGGATAACGCGGATCTGATTATTTATAACACCTGCGC >164688057-AS-4 CAGGTTGCCATACACTTTCAGTTCCGCGTTTTCACGCACCGCGCAGGTGTTATAAATAAT >164688057-SS-5 GAAAGTGTATGGCAACCTGGGCCATCTGAAACTGATTAAACGTCGTAACCCGAACCTGAA >164688057-AS-6 ACAATCGCCGGCTGCTGCATCATGCAGCCGCACACCGCAATTTTCAGGTTCGGGTTACGA >164688057-SS-7 CAGCAGCCGGCGATTGTGAAAGAAATTAAAGCGAAATATAAACATGTGGATCTGGTGTTT >164688057-AS-8 TCGCTCAGCAGTTCCGGAAATTTATACAGGTTATGGGTGCCAAACACCAGATCCACATGT >164688057-SS-9 CCGGAACTGCTGAGCGAAAGCATGAGCAGCGATAGCATTCTGATTGATGTGTGGGATGTG >164688057-AS-10 TTCGCCATCCACATCCCACACATCAATC
Primers	External >164688057FW TGCATATGAAAAGCTATTTTATTCAG >164688057RV ATCTCGAGTTCGCCATCCACATCCCA
Gene source	Gene assembly[1]

CB02	MKSYFIQTFGCQMNEHDSEKLGAMLNAMGYEPSLMADNADLIIYNTCAVRENAELKVYGNLGHLKLIKRRNPNLKIAVCGCMMQQPAIVKEIKAKYKHVDLVFGTHNLYKFPELLSESMSSLEHHHHHH
Primers	>164688057FW TGCATATGAAAAGCTATTTTATTCAG >RV_CB_v2.0 AT CTCGAG GCT GCT CAT GCT TTC GCT CAG CAG TTC
Gene source	pET-21-CB01

[1] The gene was assembled using 10 internal primers and 2 external primers. The method was adapted from (*48*).

>gi|21673629|ref|NP_661694.1| acetyl-CoA carboxylase, carboxyl transferase subunit beta/methylmalonyl-CoA mutase, C-terminus, partial [*Chlorobium tepidum* TLS]	
Full sequence	MLYSKLLADNFVCATCGHRYVRLSARDYIELILDENAFTEHQETRYIIDRDILNFPEYANKLHEERVKNGMTTALITGDGAIDGKEVVLCATSFGFLGGSFCMSTGEKVWRAAKIAIENRRPRILVAKVGQDGHDRGAKVIAAAFADIGFDVDISPLFQTPEEIVQQALDNDVHIVGISSLAGGHKTLVPQVVEGLKEARRGDILVIAGGVIPERDYDYLYERGIAGVFGPGTVIAEAAIKLLALLLEHHQ
CT01	MLYSKLLADNFVCATCGHRYVRLSARDYIELILDENAFTEHQETRYIIDRDILNFPEYANKLHEERVKNGMTTALITGDGAIDGKEVVLCATSFGFLGGSFCMSTGEKVWRAAKIAIENRRPRILVAKVGQDGHDRGAKVIAAAFADIGFDVDISPLFQTPEEIVQQALDNDVHIVGISSLAGGHKTLVPQVVEGLKEARRGDILVIAGGVIPERDYDYLYERGIAGVFGPGTVIAEAAIKLLALLLEHHQLEHHHHHH
Primers	Internal primers >21673629-SS-1 ATGCTGTATAGCAAACTGCTGGCGGATAACTTTGTGTGCGCGACCTGCGGCCATCGTTAT >21673629-AS-2 CATCCAGAATCAGTTCAATATAATCACGCGCGCTCAGACGCACATAACGATGGCCGCAGG >21673629-SS-3 TATTGAACTGATTCTGGATGAAAACGCGTTTACCGAACATCAGGAAACCCGTTATATTAT >21673629-AS-4 TTTGTTCGCATATTCCGGAAAGTTCAGAATATCACGATCAATAATATAACGGGTTTCCTG >21673629-SS-5 TCCGGAATATGCGAACAAACTGCATGAAGAACGTGTGAAAAACGGCATGACCACCGCGCT >21673629-AS-6 ACAGCACCACTTCTTTGCCATCAATCGCGCCATCGCCGGTAATCAGCGCGGTGGTCATGC >21673629-SS-7 GCAAAGAAGTGGTGCTGTGCGCGACCAGCTTTGGCTTTCTGGGCGGCAGCTTTTGCATGA >21673629-AS-8 CAATCGCAATTTTCGCCGCACGCCACACTTTTTCGCCGGTGCTCATGCAAAAGCTGCCGC
Primers	>21673629-SS-9 CGGCGAAAATTGCGATTGAAAACCGTCGTCCGCGTATTCTGGTGGCGAAAGTGGGCCAGG >21673629-AS-10 CGCAAACGCCGCCGCAATCACTTTCGCGCCACGATCATGGCCATCCTGGC

	CCACTTTCGC >21673629-SS-11 GCGGCGGCGTTTGCGGATATTGGCTTTGATGTGGATATTAGCCCGCTGTTT CAGACCCCG >21673629-AS-12 CAATATGCAC ATCGTTATCC AGCGCCTGCT GCACAATTTC TTCCGGGGTC TGAAACAGCG >21673629-SS-13 GGATAACGAT GTGCATATTG TGGGCATTAG CAGCCTGGCG GGCGGCCATA AAACCCTGGT >21673629-AS-14 TCGCCACGAC GCGCTTCTTT CAGGCCTTCC ACCACCTGCG GCACCAGGGT TTTATGGCCG >21673629-SS-15 AAGCGCGTCG TGGCGATATT CTGGTGATTG CGGGCGGCGT GATTCCGGAA CGTGATTATG >21673629-AS-16 CGGGCCAAAC ACGCCCGCAA TGCCACGTTC ATACAGATAA TCATAATCAC GTTCCGGAAT >21673629-SS-17 GGGCGTGTTT GGCCCGGGCA CCGTGATTGC GGAAGCGGCG ATTAAACTGC TGGCGCTGCT >21673629-AS-18 CTGATGATGT TCCAGCAGCA GCGCCAGCAG TTT External >21673629FW TGCATATG CTGTATAGCAAACTGCTGGC >21673629RV ATCTCGAG CTGATGATGTTCCAGCAGCAG
Gene source	Gene assembly[1]

[1] The gene was assembled using 18 internal primers and 2 external primers. The method was adapted from (*48*).

<table>
<tr><th colspan="2">>gi|153892102|ref|ZP_02013051.1| Radical SAM domain protein <i>[Opitutaceae bacterium</i> TAV2]</th></tr>
<tr><td><b>Full sequence</b></td><td>MHTHTDTNRQKPPPLSAFLPTSKKEIEARGWTDGADVILFTGDAYVDHPSFG
AAVIGRVLEAQGWRVAIVPQPNWRDDLRDFRKLGRPRLFFGISGGCMDSMV
NHYTANRRLRSDDAYTAGGMAGQRPDRVVTVYSKILKTLYPDVPLVIGGIE
ASLRRLTHYDYWSDSLRPGLLVESGADLLVYGLGEKPICEIATRLDAGEPVS
ALTDIKQTAWLATSADTAAAAGEGGGGGGETGMAGGGKSFVRHSF</td></tr>
<tr><td><b>OB01</b></td><td>MHTHTDTNRQKPPPLSAFLPTSKKEIEARGWTDGADVILFTGDAYVDHPSFG
AAVIGRVLEAQGWRVAIVPQPNWRDDLRDFRKLGRPRLFFGISGGCMDSMV
NHYTANRRLRSDDAYTAGGMAGQRPDRVVTVYSKILKTLYPDVPLVIGGIE
ASLRRLTHYDYWSDSLRPGLLVESGADLLVYGLGEKPICEIATRLDAGEPVS
ALTDIKQTAWLATSADTAAAAGEGGGGGGETGMAGGGKSFVRHSFLEHHH
HHH</td></tr>
<tr><td>Sub-cloning</td><td><i>NdeI, XhoI</i></td></tr>
<tr><td>Gene source</td><td>Plasmid 0904716_RSAM__domain_opt_pMA AmpR</td></tr>
</table>

OB02	MHTHTDTNRQKPPPLSAFLPTSKKEIEARGWTDGADVILFTGDAYVDHPSFG AAVIGRVLEAQGWRVAIVPQPNWRDDLRDFRKLGRPRLFFGISGGCMDSMV NHYTANRRLRSDDAYTAGGMAGQRPDRVVTVYSKILKTLYPDVPLVIGGIE ASLRRLTHYDYWSDSLRPGLLVESGADLLVYGLGEKPICEIATRLDAGEPVS ALTDIKQTAWLATLEHHHHHH
Primers	>SAMdom22FW CGCATATGACGAGTAAAAAAGAGATCGAAGCCCGTGGATGG >SAMdom220RV GCTCGAGTGTAGCCAGCCACGCTGTTTGTTTGATATCGG
Gene source	pET-21-OB01

>gi|15899035|ref|NP_343640.1| methylmalonyl-CoA mutase, alpha-subunit, chain B (mcmA2) [*Sulfolobus solfataricus* P2]	
Full sequence	MMITTKRIKVIVAKLGLDGHDRGAKVVARALKDAGMEVVYTGLRQTPEQIV RAALQEDADVIGISILSGAHLELIPKVVEIMKQNGLNDVGLIVGGVIPPEDIKK LKEMGVDEVFLPGSSLKEIVEKVKKVAREKRGISVE
SM01	MMITTKRIKVIVAKLGLDGHDRGAKVVARALKDAGMEVVYTGLRQTPEQIV RAALQEDADVIGISILSGAHLELIPKVVEIMKQNGLNDVGLIVGGVIPPEDIKK LKEMGVDEVFLPGSSLKEIVEKVKKVAREKRGISVELEHHHHHH
Primers	>fwMMcoANdeI CGCATATGATTACAACAAAAAGAATTAAGGTT >rvMMcoAXhoI CTCGAGTTCAACACTTATACCTCTTTT
Gene source	Genomic DNA

>gi|291284796|ref|YP_003501614.1| High-affinity branched-chain amino acid ABC transporter periplasmic leucine-specific-binding protein LivK [*Escherichia coli* O55:H7 str. CB9615]	
Full sequence	MLTHKNKATQHHEWGFLTMKRNAKTIIAGMIALAISHTALADDIKVAVVGA MSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVA VANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHI MRTAGLDSSQGPTAAKYILETVKPQRIAIIHDKQQYGEGLARSVQDGLKAAN ANVVFFDGITAGEKDFSALIARLKKENIDFVYYGGYYPEMGQMLRQARSVG LKTQFMGPEGVGNASLSNIAGDAAEGMLVTMPKRYDQDPANQGIVDALKA DKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPL NWDEKGDLKGFDFGVFQWHADGSSTKAAK
>LBP	MLTHKNKATQHHEWGFLTMKRNAKTIIAGMIALAISHTALADDIKVAVVGA MSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVA VANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHI MRTAGLDSSQGPTAAKYILETVKPQRIAIIHDKQQYGEGLARSVQDGLKAAN ANVVFFDGITAGEKDFSALIARLKKENIDFVYYGGYYPEMGQMLRQARSVG LKTQFMGPEGVGNASLSNIAGDAAEGMLVTMPKRYDQDPANQGIVDALKA DKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPL NWDEKGDLKGFDFGVFQWHADGSSTKAAKELHHHHHH
Primers	>FA_fwd ATATCGCATATGGACGATATTAAAGTCGCCGT >FC_K_rev AATCTCGAGCTTGGCCTTCGTGGATGA
Source	Genomic DNA

<table>
<tr><th colspan="2">>gi|351678212|gb|EHA61359.1| response regulator receiver protein [<i>Thermotoga maritima</i> MSB8]</th></tr>
<tr><td><b>Full sequence</b></td><td>MGKRVLIVDDAAFMRMMLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVKPFQPSRVVEALNKVSK</td></tr>
<tr><td><b>FB from CheY</b> (Fragment B for chimera building)</td><td><b>VLIVDDAAFMRMMLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVKP</b></td></tr>
<tr><td>Primers</td><td>>FB_fwd
TGAAGCCCCAGCGCGTTTTGATAGT
>FB_rev
TGGTCATAGCGGGGTTTCACAATGAA</td></tr>
<tr><td>Gene source</td><td>Plasmid pET21-CheYWT</td></tr>
<tr><td>LBP-CheY-01 (CheY inserted region in bold)</td><td>MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQR<b>VLIVDDAAFMRMMLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVKP</b>RYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKGFDFGVFQWHADGSSTKAKELHHHHHH</td></tr>
<tr><td>Primers</td><td>>FA_fwd
ATATCGCATATGGACGATATTAAAGTCGCCGT
>FA_rev
ACTATCAAAACGCGCTGGGGCTTCA
>FB_fwd
TGAAGCCCCAGCGCGTTTTGATAGT
>FB_rev
TGGTCATAGCGGGGTTTCACAATGAA
>FC_fwd
TTCATTGTGAAACCCCGCTATGACCA
>FC_K_rev
AATCTCGAGCTTGGCCTTCGTGGATGA</td></tr>
<tr><td>Gene source</td><td>Plasmids pET21-CheYWT and pET21-LBPWT</td></tr>
<tr><td>LBP-CheY-02 (CheY inserted region in bold)</td><td>MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQR<b>VLIVDDAAFMRMMLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVKP</b>KRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKGFDFGVFQWHADGSSTKAKELHHHHHH</td></tr>
<tr><td></td><td>>FA_fwd
ATATCGCATATGGACGATATTAAAGTCGCCGT
>FA_rev
ACTATCAAAACGCGCTGGGGCTTCA
>FB_fwd
TGAAGCCCCAGCGCGTTTTGATAGT
>FB_K_rev
TGGTCATAGCGTTTGGGTTTCACAATGAA
>FC_K_fwd
TTCATTGTGAAACCCAAACGCTATGACCA
>FC_K_rev
AATCTCGAGCTTGGCCTTCGTGGATGA</td></tr>
<tr><td>Gene source</td><td>Plasmids pET21-CheYWT and pET21-LBPWT</td></tr>
</table>

>gi|351678212|gb|EHA61359.1| response regulator receiver protein [*Thermotoga maritima* MSB8]	
LBP-CheY-03 (CheY inserted region in bold)	MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQR**VLIVDDAAFMRMMLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDFIVKPK**RYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKGFDFGVFQWHADGSSTKAK
Primers	>FA_fwd ATATCGCATATGGACGATATTAAAGTCGCCGT >chimera-lbp-chey-c-ter-A344>K AATC TCGA GTCA CTTG GGCT TCGT GGAT GA
Gene source	Plasmid pET21-LBP-CheY-02
LBP-CheY-04 (CheY inserted region in bold)	MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQR**ILIVDDAAFMRMMLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAKDF**VTMPKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKGFDFGVFQWHADGSSTKAK
Primers	>FA_fwd ATATCGCATATGGACGATATTAAAGTCGCCGT >OT01_B_I_fwd TGAAGCCCCAGCGCAUATTGATAGTcgatga >OT01_B_MTV_rev TGGTCATAGCGTTTGGGCATCGTAACGAAGTCTTT >OT01_A_I_rev tcatcgACTATCAAtAtGCGCTGGGGCTTCA >OT01_C_MTV_fwd AAAGACTTCGTTACGATGCCCAAACGCTATGACCA >chimera-lbp-chey-c-ter-A344>K AATC TCGA GTCA CTTG GGCT TCGT GGAT GA
Gene source	Plasmid pET21-LBP-CheY-02
LBP-CheY-04 (CheY inserted region in bold)	MDDIKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPTAAKYILETVKPQR**VLIVDDAAFMRMMLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMNGIDAIKEIMKIDPNAKIIV**GPEGVGNASLSNIAGDAAEGMLVTMPKRYDQDPANQGIVDALKADKKDPSGPYVWITYAAVQSLATALERTGSDEPLALVKDLKANGANTVIGPLNWDEKGDLKGFDFGVFQWHADGSSTKAK
Primers	>FA_fwd ATATCGCATATGGACGATATTAAAGTCGCCGT >RV_mid-LBPCHEY_version10 ACC TTC CGG CCC GAC GAT GAT CTT >FW_mid-LBPCHEY_version10 AAG ATC ATC GTC GGG CCG GAA GGT >chimera-lbp-chey-c-ter-A344>K_rev_stop _TCA AAT CTC GAG *TCA* CTT GGC CTT CGT GGA TGA
Gene source	Plasmid pET21-LBP-CheY-03

8.4 Full table of intermediate sequences.

The headers of the columns designate the following information:

GI: Gene identifier from intermediate sequence
Protein name: Automatic annotation from UniProt database
Source and Taxonomy: Taxonomy of source organism

c.23: SCOP and ASTRAL identifiers for flavodoxin-like fold structures compared with the intermediate sequence
cols: Number of HMM-HMM aligned columns between intermediate sequence and the structures from flavodoxin-like and $(\beta\alpha)_8$-barrel fold structures
c.1: SCOP and ASTRAL identifiers for $(\beta\alpha)_8$-barrel fold structures compared with the intermediate sequence
Id.%: Percentage of identity between the intermediate sequence and the sequence of the compared structure

Table 23: Full list of intermediate candidates

GI	Protein name	Source and taxonomy	c.23	cols	Id.%	c.1	cols	Id.%
150401229	radical SAM domain-containing protein	*Methanococcus aeolicus* Archaea; Euryarchaeota	d1reqa2 c.23.6.1	128	0.17	d1ujpa_ c.1.2.4	101	0.18
117619094	response regulator	*Aeromonas hydrophila* Bacteria; Proteobacteria; Gammaproteobacteria	d1p6qa_ c.23.1	109	0.17	e2basa1 c.1.33	245	0.23
54026608	putative two-component response regulator	*Nocardia farcinica* Bacteria; Actinobacteria	d1p6qa_ c.23.1	118	0.31	d1o4ua1 c.1.17	97	0.14
46202454	COG2185: Methylmalonyl-CoA mutase	*Magnetospirillum* Bacteria; Proteobacteria; Alphaproteobacteria	d1reqa2 c.23.6.1	134	0.27	d1vc4a_ c.1.2.4	84	0.21
116624963	radical SAM domain-containing protein	*Solibacter usitatus* Bacteria; Fibrobacteres	d1reqa2 c.23.6.1	74	0.23	d1vc4a_ c.1.2.4	97	0.18
149914223	cobalamin B12-binding protein	*Roseobacter sp* Bacteria; Proteobacteria; Alphaproteobacteria	d1reqa2 c.23.6.1	143	0.16	d1y0ea_ c.1.2.5	103	0.13
162453798	hypothetical protein sce5522	*Sorangium cellulosum* Bacteria; Proteobacteria; delta/epsilon subdivisions	d1reqa2 c.23.6.1	100	0.19	d1pii_1 c.1.2.4	72	0.17
83746130	Sensory transduction protein kinase	*Ralstonia solanacearum* Bacteria; Proteobacteria; Betaproteobacteria	d1p6qa_ c.23.1.1	109	0.17	e2basa1 c.1.33	245	0.21
153892102	Radical SAM domain protein	*Opitutaceae bacterium* Verrucomicrobia	d1reqa2 c.23.6.1	110	0.25	d1o94a1 c.1.4.1	86	0.1
83859913	sigma-54 dependent DNA-binding response regulator	*Oceanicaulis alexandrii* Bacteria; Proteobacteria; Alphaproteobacteria	d1ny5a1 c.23.1.1	132	0.36	d1o4ua1 c.1.17.1	93	0.19
86740508	radical SAM family protein	*Frankia sp. CcI3* Bacteria; Actinobacteria; Actinobacteria	d1reqa2 c.23.6.1	135	0.17	d1vyra_ c.1.4.1	93	0.14
137771362	hypothetical protein GOS_6831935	*marine metagenome*	d1reqa2 c.23.6.1	88	0.38	e1yxya1 c.1.2.5	83	0.19
140978759	hypothetical protein GOS_3993934	*marine metagenome*	d1reqa2 c.23.6.1	149	0.17	d1twda_ c.1.30	223	0.36
8675109	cobalamin B12-binding	*Rhodopseudomonas palustris* Bacteria; Proteobacteria; Alphaproteobacteria	d1reqa2 c.23.6.1	147	0.21	d2tpsa_ c.1.3.1	96	0.25

GI	Protein name	Taxonomy	c.23	cols	Id.%	c.1	cols	Id.%
156937162	Radical SAM domain-containing protein	*Ignicoccus hospitalis* Archaea; Crenarchaeota; Thermoprotei	e1ys7a2 c.23.1.1	111	0.21	d1a53__ c.1.2.4	89	0.2
163753828	Fe-S protein, radical SAM family	*Kordia algicida OT-1* Bacteroidetes/Chlorobi group	d1reqa2 c.23.6.1	89	0.16	d1mxsa_ c.1.10.1	80	0.13
33359517	btpA family protein	*Pyrococcus furiosus* Archaea; Euryarchaeota; Thermococci	d1ccwa_ c.23.6.1	103	0.16	d1geqa_ c.1.2.4	177	0.2
15899035	methylmalonyl-CoA mutase	*Sulfolobus solfataricus* Archaea; Crenarchaeota; Thermoprotei	d1reqa2 c.23.6.1	132	0.39	e1yxya1 c.1.2.99	122	0.26
20094042	hypothetical protein MK0604	*Methanopyrus kandleri* Archaea; Euryarchaeota; Methanopyri	d1reqa2 c.23.6.1	64	0.19	d1vhna_ c.1.4.1	211	0.23
143100560	hypothetical protein GOS_1409136	*marine metagenome*	d1reqa2 c.23.6.1	123	0.11	d1twda_ c.1.30.1	238	0.38
134467176	hypothetical protein GOS_220031	*marine metagenome*	d1reqa2 c.23.6.1	122	0.16	d1rd5a_ c.1.2.4	114	0.15
168701434	CutC family protein	*Gemmata obscuriglobus* Bacteria; Planctomycetes; Planctomycetacia	d1reqa2 c.23.6.1	149	0.16	d1twda_ c.1.30	234	0.33
144006038	hypothetical protein GOS_472759	*marine metagenome*	d1reqa2 c.23.6.1	89	0.19	d1pii_1 c.1.2.4	253	0.37
149173777	acid aldolase protein	*Planctomyces maris* Bacteria; Planctomycetes; Planctomycetacia	d1a04a2 c.23.1.1	104	0.12	d1dxea_ c.1.12	248	0.32
146319339	response regulator	*Streptococcus suis* Bacteria; Firmicutes; Bacilli	d1a04a2 c.23.1.1	134	0.31	d1qpoa1 c.1.17.1	95	0.15
134299625	cobalamin B12-binding domain-containing protein	*Desulfotomaculum reducens* Bacteria; Firmicutes; Clostridia	d1reqa2 c.23.6.1	134	0.33	e1tv5a1 c.1.4.1	103	0.19
119477187	response regulator	*marine gamma* Bacteria; Proteobacteria; Gammaproteobacteria	d1k66a_ c.23.1.1	110	0.17	e2basa1 c.1.33	246	0.19
51948340	APRR Response regulator plant-like protein	*Ostreococcus tauri* Eukaryota; Viridiplantae; Chlorophyta	d1i3ca_ c.23.1.1	141	0.21	e2basa1 c.1.33	122	0.13
188991972	hypothetical protein xccb100_2577	*Xanthomonas campestris* Bacteria; Proteobacteria; Gammaproteobacteria	d1reqa2 c.23.6.1	110	0.16	d1vc4a_ c.1.2.4	92	0.21
51245292	hypothetical protein DP1440	*Desulfotalea psychrophila* Bacteria; Proteobacteria; delta/epsilon subdivisions	d1qkka_ c.23.1.1	139	0.17	d1o4ua1 c.1.17.1	96	0.15
114777666	hypothetical protein SPV1_08361	*Mariprofundus ferrooxydans* Bacteria; Proteobacteria;	d1mb3a_ c.23.1.1	108	0.15	e2basa1 c.1.33	247	0.2
135504475	hypothetical protein GOS_9154081	*marine metagenome*	d1reqa2 c.23.6	108	0.18	d1qopa_ c.1.2.4	95	0.12
135387500	hypothetical protein GOS_9286769	*marine metagenome*	d1reqa2 c.23.6	100	0.31	d2tpsa_ c.1.3.1	93	0.26
118050939	response regulator receiver modulated diguanylate phosphodiesterase	*Comamonas testosteroni* Bacteria; Proteobacteria; Betaproteobacteria	d1mb3a_ c.23.1.1	114	0.15	e2basa1 c.1.33	244	0.18
136705953	hypothetical protein GOS_7841535	*marine metagenome*	d1reqa2 c.23.6.1	167	0.49	e1yxya1 c.1.2.5	134	0.12
39934610	coenzyme B12-binding	*Rhodopseudomonas* Bacteria; Proteobacteria; Alphaproteobacteria	d1reqa2 c.23.6.1	147	0.2	d1geqa_ c.1.3.1	83	0.18
34495765	hypothetical protein CV_0310	*Chromobacterium violaceum* Betaproteobacteria	d1reqa2 c.23.6.1	118	0.14	e2basa1 c.1.33	171	0.11

GI	Protein name	Taxonomy	c.23	cols	Id.%	c.1	cols	Id.%
77463853	regulatory protein, PpaA	*Rhodobacter sphaeroides* Bacteria; Proteobacteria; Alphaproteobacteria	d1reqa2 c.23.6.1	146	0.16	d1y0ea_ c.1.2.5	103	0.12
23015349	COG1032: Fe-S oxidoreductase	*Magnetospirillum* Bacteria; Proteobacteria; Alphaproteobacteria	d1reqa2 c.23.6.1	126	0.15	d1i4na_ c.1.2.4	102	0.11
108800290	cobalamin B12-binding protein	*Mycobacterium sp.* Bacteria; Actinobacteria; Actinobacteria (class)	d1reqa2 c.23.6.1	148	0.16	d1y0ea_ c.1.2.5	102	0.13
119695507	cobalamin B12-binding domain protein	*Mycobacterium sp* Bacteria; Actinobacteria; Actinobacteria (class)	d1reqa2 c.23.6.1	148	0.16	d1y0ea_ c.1.2.5	102	0.13
73749252	DNA-binding response regulator	*Dehalococcoides sp.* Bacteria; Chloroflexi; Dehalococcoidete	d1mvoa_ c.23.1.1	116	0.35	d1rd5a_ c.1.2.4	108	0.18
116750679	cobalamin B12-binding domain-containing protein	Syntrophobacter fumaroxidans Bacteria; Proteobacteria;	d1reqa2 c.23.6.1	127	0.18	d1rd5a_ c.1.2.4	101	0.17
18976592	hexulose-6-phosphate synthase	Pyrococcus furiosus Archaea; Euryarchaeota; Thermococci	d1kgsa2 c.23.1.1	112	0.15	d1rd5a_ c.1.2.4	246	0.18
150391225	radical SAM domain-containing protein	Alkaliphilus metalliredigens Bacteria; Firmicutes; Clostridia; Clostridiales	d1reqa2 c.23.6.1	122	0.16	d1hg3a_ c.1.1.1	102	0.15
163854214	cobalamin B12-binding	Methylobacterium Bacteria; Proteobacteria; Alphaproteobacteria	d1bmta2 c.23.6.1	129	0.14	d1y0ea_ c.1.2.5	118	0.14
157374162	response regulator receiver modulated diguanylate phosphodiesterase	Shewanella sediminis Bacteria; Proteobacteria; Gammaproteobacteria	d1k66a_ c.23.1.1	109	0.15	e2basa1 c.1.33	246	0.25
135280051	hypothetical protein GOS_9405829	marine metagenome	d1jbea_ c.23.1.1	101	0.22	e2c0aa_ c.1.10.1	209	0.47
140452724	hypothetical protein GOS_5212998	marine metagenome	d1ccwa_ c.23.6.1	94	0.31	e1yxya1 c.1.2.5	90	0.16
116751500	radical SAM domain-containing protein	Syntrophobacter Bacteria; Proteobacteria; delta/epsilon subdivisions	d1reqa2 c.23.6.1	123	0.21	d1vyra_ c.1.4.1	85	0.14
152964427	response regulator receiver modulated FAD-dependent pyridine nucleotide-disulphide oxidoreductase	Bacteria; Actinobacteria; Actinobacteria (class)	d1k68a_ c.23.1.1	135	0.19	e2c0aa_ c.1.10.1	87	0.15
143306820	hypothetical protein GOS_1166505	marine metagenome	d1reqa2 c.23.6.1	109	0.17	d1mo0a_ c.1.1.1	93	0.15
126180333	radical SAM domain-containing protein	Methanoculleus marisnigri Archaea; Euryarchaeota; Methanomicrobia	d1ccwa_ c.23.6.1	107	0.15	d1vyra_ c.1.1.4	99	0.16
31335370	glutamate mutase subunit A	Actinoplanes friuliensis Bacteria; Actinobacteria; Actinobacteria (class)	d1ccwa_ c.23.6.1	131	0.22	d1rd5a_ c.1.2.4	87	0.23
163798434	Radical SAM domain protein	Methanococcus voltae Archaea; Euryarchaeota; Methanococci	d1reqa2 c.23.6.1	121	0.13	d1ujpa_ c.1.2.4	134	0.14
152982555	hypothetical protein mma_0744	Janthinobacterium sp Bacteria; Proteobacteria; Betaproteobacteria	d1reqa2 c.23.6.1	111	0.15	e2basa1 c.1.33	203	0.18
15606339	nicotinate phosphoribosyltransferase frag	Aquifex aeolicus Bacteria; Aquificae; Aquificae (class)	d1reqa2 c.23.6.1	72	0.32	e1ytka1 c.1.17	74	0.43
33592924	regulatory protein BvgR	Bordetella pertussis Bacteria; Proteobacteria; Betaproteobacteria	d1i3ca_ c.23.1.1	107	0.18	e2basa1 c.1.33	230	0.11
143697032	hypothetical protein GOS_800004	marine metagenome	d1ccwa_ c.23.6.1	126	0.27	d1gqna_ c.1.10.1		0.13
142146541	hypothetical protein GOS_2809782	marine metagenome	d1reqa2 c.23.6.1	107	0.3	d1zfja1 c.1.5.1	81	0.17

GI	Protein name	Taxonomy	c.23	cols	Id. %	c.1	cols	Id. %
135253658	hypothetical protein GOS_9431870	marine metagenome	d1reqa2 c.23.6.1	157	0.23	e1yxya1 c.1.2.5	105	0.17
163698961	Methylmalonyl-CoA mutase	Methylobacterium nodulans Bacteria; Proteobacteria; Alphaproteobacteria	d1reqa2 c.23.6.1	166	0.33	e1yxya1 c.1.2.4	136	0.2
138518049	hypothetical protein GOS_6026896	marine metagenome	d1reqa2 c.23.6.1	164	0.29	d1rd5a_ c.1.2.4	156	0.19
138632307	hypothetical protein GOS_6362113	marine metagenome	d1reqa2 c.23.6.1	89	0.43	e1yxya1 c.1.2.5	83	0.17
136840365	hypothetical protein GOS_7702467	marine metagenome	d1reqa2 c.23.6.1	125	0.32	d1y0ea_ c.1.2.5	120	0.17
147678754	methylaspartate mutase subunit S	Pelotomaculum thermopropionicum	d1reqa2 c.23.6.1	134	0.48	d1y0ea_	c.1.2.5	0.12

8.5 Acknowledgments

To Guadalupe, Juan Jose, Leon Fernando and Victor Hugo, my fabulous family. Without their encouragement and support I could have not achieved anything.

I wish to thank Birte Höcker for all the support and advice she always offered to me.

I want to acknowledge the scientific advice from Prof. V. Braun and Prof. A. Lupas

For our great working environment I want to thank all my lab colleagues: Sooruban, Kaspar, Simone, Nils, Andre, Bettina, Chris, Bruce, Simon, Dominik and Sara. Not only their scientific contributions were determinant for my research. Also the funny moments and interesting meals were inspiring.

I want to specially thank Saacnicteh because all her support in challenging moments of my work. I really enjoyed working with my friend and colleague Ulrike.

I am particularly grateful for the help given by Vikram and Moritz. Their advice in my first years working at the institute was very significant.

I am also grateful to the crystallography community at the MPI for their support, especially Fulvia Bono, Marcus Hartmann, and Oliver Weichenrieder. Without their help my work would not have been possible.

I would like to express my gratitude to Lorenzo Segovia, because sparking my interest in the fascinating world of the proteins

For sharing inspiring experiences, which encouraged me to continue with my work, I want to thank my friends Mike and Marco. For our delightful scientific conversations I want to thank my friends Andreas and Rodrigo.