UNIVERSITÄT BAYREUTH

"Investigations into the evolution of a periplasmic binding protein and its implications in the origins of modern protein folds"

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
an der Fakultät für Biologie, Chemie und Geowissenschaften
der Universität Bayreuth

vorgelegt von

Florian Michel

aus Treuchtlingen

Bayreuth, 2023

Die vorliegende Arbeit wurde in der Zeit von Dezember 2017 bis Dezember 2023 in Bayreuth am Lehrstuhl Biochemie III unter Betreuung von Frau Prof. Dr. Birte Höcker angefertigt.

Vollständiger Abdruck der von der Fakultät für Biologie, Chemie und Geowissenschaften an der Universität Bayreuth genehmigten Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. Nat.)

Art der Dissertation: Kumulative Dissertation

Dissertation eingereicht am: 12.12.2023

Zulassung durch die Promotionskommission: 20.12.2023

Wissenschaftliches Kolloquium: 22.04.2024

Amtierender Dekan: Prof. Dr. Cyrus Samimi

Prüfungsausschuss:

Prof. Dr. Birte Höcker	(Gutachterin)
Prof. Dr. Ulrich Krauß	(Gutachter)
Prof. Dr. Carlo Unverzagt	(Vorsitz)
PD Dr. Alfons Weig	

“The universe is, instant by instant, recreated anew. There is in truth no past, only a memory of the past. Blink your eyes, and the world you see next did not exist when you closed them. Therefore, the only appropriate state of the mind is surprise. The only appropriate state of the heart is joy. The sky you see now, you have never seen before.

The perfect moment is now. Be glad of it.”

- Terry Pratchett -

Contents

Zusammenfassung

Biomakromoleküle als Bausteine des Lebens sind verantwortlich für den Ablauf fast aller biologischen Prozesse. Neben Nukleinsäuren, Kohlenwasserstoffen und Lipiden spielt die Gruppe der Proteine hierbei eine zentrale Rolle.

Wenn wir bis ins Detail verstehen wollen, wie die Natur diese Vielzahl an Funktionen hervorgebracht hat bietet es sich an, die evolutionäre Geschichte der Proteine näher unter die Lupe zu nehmen. Durch Erforschung der molekularen Prozesse, die bei der Evolution von Proteinen eine Rolle spielten, lernen wir nicht nur die Grundregeln wie sich die Struktur von natürlichen Proteinen aufbaut, sondern können dieses Wissen auch in Zukunft für unsere Zwecke verwenden.

Durch den Aufstieg von immer sensibleren Methoden der Sequenz- und Strukturanalyse können wir eine davor unerahnte Menge an Information für gezielte Forschungszwecke nutzen (Paper V). Systematische Analyse der Regeln und strukturellen Gegebenheiten der frühen Proteinevolution bieten Einsicht in eben diese grundlegenden Spielregeln. In dieser Arbeit soll am Beispiel eines periplasmatischen Bindeproteins (PBP) dieser Weg von kleinen, subdomänengroßen Struktureinheiten – den grundlegenden Bausteinen – über vorhergegangene, schon komplexere Proteinstrukturen bis hin zu der Form, die wir heute in der Natur beobachten, verfolgt werden. Zu diesem Zweck wurden, mithilfe der Datenbank *Fuzzle* subdomänengroße Fragmente in PBPs identifiziert (Paper III). Eines dieser Proteine, das Ribose-Bindeprotein von *T. maritima* (RBP) wurde auf Basis dieser Analyse als Modelsystem im Labor erforscht. Die strukturelle Analyse der aus dem RBP isolierten Fragmente zeigt, dass diese Bausteine auch dann eine grundlegende Stabilität aufweisen, wenn sie aus dem Kontext des parentalen Proteins entnommen werden (Paper IV). Durch spätere Anlagerung oder Duplikation weiterer Elemente erreichen wir in diesem Gedankenspiel nun eine mögliche Vorstufe eines Proteins mit einer den Flavodoxinen ähnelnden Struktur. Die Flavodoxin-ähnliche Faltung gilt als ein Vorgänger der modernen PBP. Durch eine Duplikation dieses als Vorgänger geltenden Proteins und der Aneignung der Funktion, spezifische Liganden zu binden, erreichen wir letztendlich das moderne PBP mit seiner charakteristischen, symmetrischen Struktur mit einer Bindetasche zwischen den sich gegenüberstehenden Einzeldomänen.

Um die jeweiligen Einzeldomänen des konkreten Beispiels des RBPs zu isolieren, wurden zyklische Permutationen der N- und C-terminalen Domänen generiert (Paper II). Eine strukturelle Analyse dieser jetzt als vollständige Domänen geltenden Strukturen zeigt, dass sie sich im Kern mit der als einen potenziellen strukturellen Vorgänger vermuteten Flavodoxin-ähnlichen Faltung vergleichen lassen.

In einer weiteren Studie wurden zusätzlich auch die beiden linear „zerschnittenen“ Einzeldomänen analysiert (Paper I). Es zeigt sich, dass sich die beiden Domänen in einer definierten Struktur wiederfinden lassen. Bei Coexpression der beiden Domänen bilden diese eine Heterodimer, bei gleichzeitiger Rekonstitution der Funktion des RBPs, Ribose zu binden.

Durch eine systematische Analyse von weiteren Proteinen ähnlich wie in dieser Arbeit könnten wir nicht nur unser Verständnis von Proteinfaltung im Einzelnen erweitern, sondern auch unser Wissen über die Prozesse des frühen Lebens ausbauen. Auch die hohe Modularität des Modellsystems könnte nützlich sein, um weitere Einsatzmöglichkeiten von PBPs in Forschung und Technik zu entwickeln.

Abstract

As the building blocks of life, biomacromolecules are responsible for almost all biological processes. Alongside nucleic acids, hydrocarbons and lipids, proteins play a central role.

If we want to understand in detail how nature has produced this multitude of functions, it is worth taking a closer look at the evolutionary history of proteins. By studying the molecular processes that played a role in the evolution of proteins, we not only learn the basic rules of how the structure of natural proteins is built up, but can also use this knowledge for our purposes in the future.

With the rise of increasingly sensitive methods of sequence and structural analysis, we can utilize a previously inaccessible amount of information (Paper V). Systematic analysis of the rules and structural features of early protein evolution provides insight into these fundamental processes. In this work, we will use the example of a periplasmic binding protein (PBP) to trace this path from small subdomain units - the basic building blocks – via more complex progenitor protein structures to the structure which we can observe in nature today. To this end, subdomain-sized fragments in PBPs were identified using the *Fuzzle* database (Paper III). One of these proteins, the ribose-binding protein of *T. maritima* (RBP), was investigated as a model system in the laboratory based on this analysis. Structural analysis of the fragments isolated from RBP shows that these building blocks exhibit fundamental stability even when removed from the context of the parental protein (Paper IV). By accreting or duplicating further elements, we can now build a possible precursor of a protein with a structure similar to that of flavodoxins – a fold thought to be a precursor of PBPs. By duplicating this protein and subsequent functionalization of binding specific ligands, we ultimately obtain the modern PBP with its characteristic symmetrical structure with a binding pocket between the opposing individual domains.

To now isolate the respective single domains of the specific example of the RBP, cyclic permutations of the N- and C-terminal domains were generated (Paper II). A structural analysis of these now complete domains shows that they can be compared in their core with the flavodoxin-like fold, which is thought to be a potential structural predecessor.

In a further study, the two linearly "cut" single domains were also analyzed (Paper I). It was shown that the two domains maintain their defined structure. When the two domains are co-expressed, they form a heterodimer, while simultaneously reconstituting the function of the RBP to bind ribose.

By expanding and systematically analyzing other proteins similar to this work, we could not only expand our understanding of protein folding in detail, but also expand our knowledge of the processes of early life. The high modularity of the model system could also be used to explore further applications of PBPs in research and technology applications.

1. Introduction

1.1. The stability and folding of natural proteins – An enigma

For more than half a century the central dogmas of protein folding have more or less stood firm: The central idea that the flow of information happens from nucleic acid to proteins, the Anfinsen dogma that proteins only fold into one, native conformation and the Levinthal's paradox that proteins do not sample the entire possible three-dimensional space on their transition to their native structure[1,2,3,4]. While these results have shown for the first time that a polypeptide chain of a certain sequence can spontaneously adopt its natural conformation *in vivo*, this process for any given protein is still a mystery to us, with many aspects left to discover.

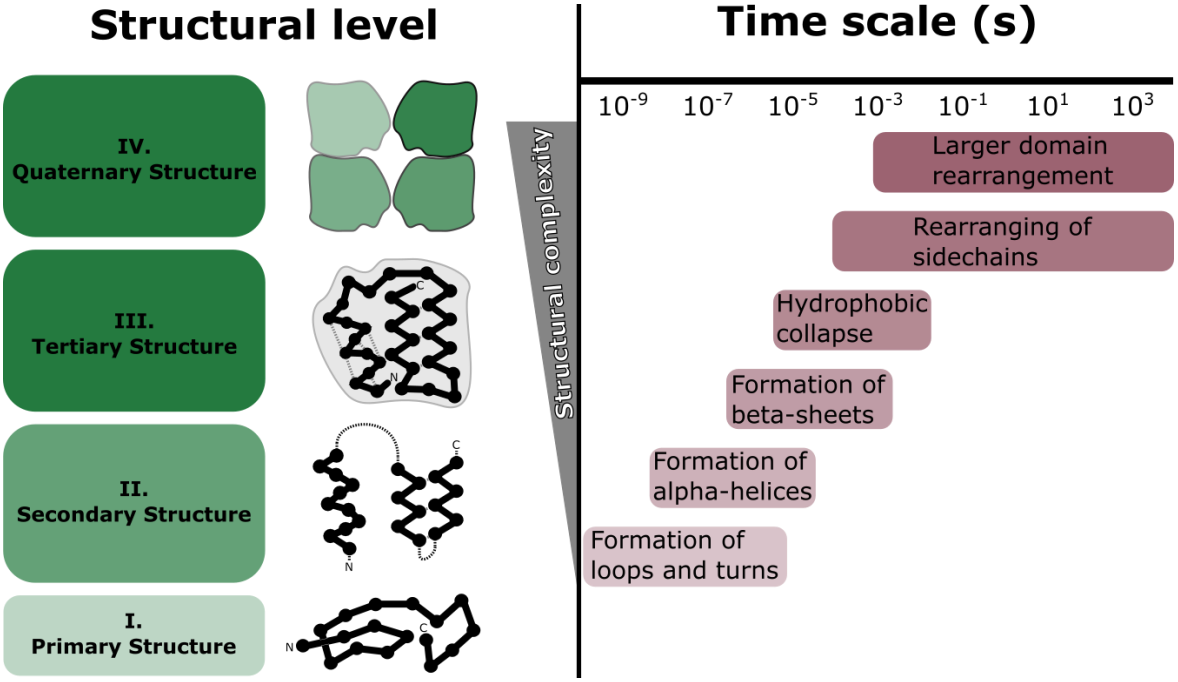


Figure 1: Schematic representation of the structural hierarchy and timescales at play during protein folding. The main structural levels during protein folding and their rise in structural complexity, from a single unstructured poly-peptide chain to a folded, globular protein (left). Beginning from the primary, unfolded polypeptide chain (I), the developing of secondary structure elements (II), formation of tertiary structure (III) and assembly into a quaternary structure (IV). Time scales associated with common events during protein folding (left).

Since unfolded, denatured polypeptide chains do not only interact with the surrounding environment, but also have many interactions within the molecule itself, folding is a complex process. Generally, the folding of a protein can be broken down into several major steps, that must happen in sequence to reach the native conformation (Figure **1**). The first step is the exchange of external solvent interactions in favor of internal

interactions, with hydrophobic interactions being the main driver of this process[5]. This step happens extremely fast, limited only by the diffusion speed of the polypeptide chain folding onto itself. In these first nano- to microseconds, first regions of the chain are already forming a rudimentary secondary structure of the protein[6,7].

As a consequence of this initial collapse, amino acids with hydrophobic side chains begin to become buried in a now solvent-excluded interior, while polar and charged side-chains favor interactions with the outside solvent. This leads to what is generally termed the hydrophobic collapse of the protein. Since investigation of these fast early events in folding is a challenge, it is still unclear whether this step is universal in the folding pathway of globular proteins[8,9,10]. Some studies show that hydrophobic collapse might occur before the formation of secondary structure[11]. Other studies of several model proteins used in the investigation of protein folding such as lysozyme[12,13], triosephosphate isomerase[14], barstar[15,16,17], ribonuclease[2,18,] or myoglobin[19,20] have shown that the hydrophobic collapse is an integral step in the folding of these proteins. After the formation of this more compact structure, the next crucial step is the rearrangement of sidechains. Native sidechain orientation can then be achieved via sidechain-sidechain interaction, interactions with the protein backbone, and rearrangement via larger domain movements. These last steps that happen on a microsecond to second scale are the last on the path to obtaining the definite, precisely formed native conformation any given protein needs to be able to carry out its function[21].

Other models of the early stages of folding include the *framework* or *diffusion-collision model*, the *nucleation model,* and the model of hydrophobic collapse. The *diffusion-collision model* presumes that local stretches of native-like secondary structure form independently as a basis for the correct formation of the tertiary structure[20,22,23]. These partially formed elements then diffuse until they come in contact with other corresponding elements and fold into their native tertiary structure. Similarly, the *nucleation model* first presumes the formation of a native-like secondary structure based on interactions of neighboring side chains and the backbone, but reaches the tertiary structure as a consequence of this previous nucleation of the secondary structure through specific interactions[24]. Both the *diffusion-collision* and the *nucleation model* propose that secondary structure is mostly formed prior to tertiary structure, whereas in the model of hydrophobic collapse the formation of secondary and tertiary structure is not proceeding sequentially during folding[25,10,9].

All these mechanisms and interactions however offer only relatively small individual contributions to the total stabilizing energy of a given protein (Figure 2). Only in combination can the stabilizing effects of hydrophobic, entropic and enthalpic interactions overcome the destabilizing terms of chain entropy and hydrophilic interactions with the solvent[26,27]. These underlying circumstances govern the properties of all proteins[28]. As a consequence, the precisely set equilibrium of destabilizing and stabilizing energies lead to proteins being only marginally stable[29,30], meaning that the net contribution of energy stabilizing the protein (ΔG^0) is relatively small compared to the other energies at play.

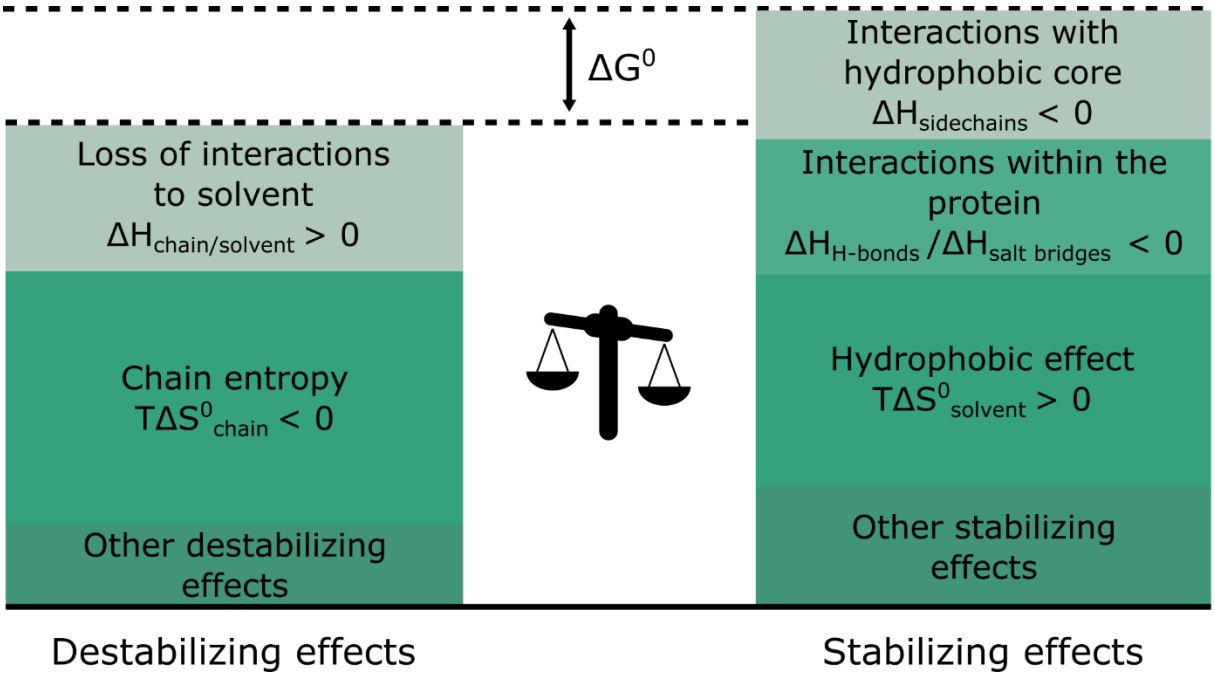


Figure 2:Overview of stabilizing and destabilizing effects that govern total protein stability. Enthalpic and entropic energy terms and their contribution to either destabilization (left column) and stabilization (right column) of folded protein structures. The difference in energy (ΔG^0) represents the total free energy of the protein and is formed by a relatively small offshoot in the equilibrium of the many effects at play.

This circumstance might seem to contain a paradox on first sight; however, it provides crucial advantages. Not only does the marginal stability of proteins allow for more inherent flexibility of their structure, and thus for example important side-chain movements needed for the function of the protein[31], but also expands the sequence space accessible for evolution. While proteins with high stabilities could compensate substitutions necessary for the emergence of new functions more easily, proteins with lower structural stability offer access to “bridge states” in the sequence space. These “bridge states” are regions in the sequence space in which different structures overlap.

These interstitial spaces could allow for new evolutionary trajectories via these “bridge states”, which are inaccessible to proteins of high stability[30].

The marginal stability ensures a sufficient turn-over in a cell environment, since a protein with an energy barrier too high to allow for degradation in a physiological context would accumulate over the life cycle of an organism[32,33]. It has been hypothesized that transition states in folding are not universally necessary and could be a mechanism to ensure cooperativity of folding and stabilize the native state[34]. Colloquially one could propose that a protein is ever just as stable as it needs to be to carry out its function, which is also reflected by the lack of evolutionary pressure for higher stabilities. This corresponds to the behavior of the first computer-generated proteins, which were designed to be as stable as possible and with a focus on optimizing folding rates, showing a much higher stability than natural proteins[35].

These circumstances in protein folding when observed in natural proteins are one of the main reasons why their investigation is so important. This balancing act of energy terms that lead to a stable, yet flexible protein is something that is generally not in the scope of protein design approaches. Generally, computer-based approaches aim to minimize the energy state of a protein, potentially disfavoring marginally stable solutions. This often leads to computer-designed proteins being extremely stable, and thermodynamic stability being one of the main design goals[36,37,38]. While recent advances in this field can reliably produce very stable scaffolds for a variety of binding functions, something we still struggle with is the design of enzymatic functions. Furthering our knowledge of the marginal stability of proteins could one day help us learn how to also recapture this feature in designs[39].

1.2. Folding in the context of Protein Evolution – Why it matters

To bolster our knowledge on how protein folding works in general and how to leverage this knowledge for protein design, retracing the steps of protein evolution could offer valuable insights. To investigate the origin of life on our planet however is a complicated matter. There is no way to directly go into the past and investigate, and e. Evolution is an unstopping process, ever changing the status-quo of life. The science of evolution, the origin of life and its path to the modern day is somewhat akin to how

humanity has treated knowledge for millennia: By telling stories, a painting on the wall of a cave, writing down fantastic tales, singing songs about a successful hunt, we are passing down knowledge from one generation to the next, deviating and adapting to new circumstances. Similar to how linguists operate when trying to figure out the etymology of a word, we can apply the same methods for the investigation of molecular evolution[40,41]. Whereas the challenge for a linguist is to correctly infer the historical context of a word by tracing the change of the word through the course of time, a scientist investigating the molecular evolution of a protein can implement similar methods by tracing the changes in the protein sequence[42]. This is also reflected in the vocabulary shared to convey concepts both in linguistics and evolutionary sciences, for example the idea that the origin of a certain word/protein could stem from the same origin (homologues) or happen to share similarities while not sharing a common origin (analogues). Similarly, this concept has been picked up by John Maynard Smith in his 1970s article about traversing the protein space, creating an analogy that is still popular today[43,44].

While the protein world we can observe today is incredibly sophisticated, it is the result of billions of years of evolution. However, the mechanism of evolution did not have to reinvent the groundwork of this system for each new protein. It has used a powerful machinery of processes that led to new protein structures, added function, and refined it[45]. The make-or-break process in the world of proteins is its transformation from an unstructured string of amino acids into a three-dimensional structure[46]. Ultimately, the given sequence of amino acids governs this shape, and determines the role of a protein.

Completely understanding how this three-dimensional puzzle works would help to apply this knowledge for our own advantage, such as helping us create new proteins with tailor-made functions, create new design tools, engineer protein building blocks or invent complex multi-protein systems[47,48,49,50]. The idea that during the evolution of a modern protein the original amino acid sequence could not have sampled the entirety of the theoretical space of conformations is described in the famous Levinthal's paradox[3]. If the sampling of the conformational space happened with equal probability, a polypeptide chain cannot spontaneously fold in the timeframes we observe in nature. The logical explanation to this 'paradox' at first glance is rather simple: Protein folding is facilitated by a rapid, energetically favored formation of local interactions which nudge the fold into a certain conformation, a concept also first described by Levinthal

in 1968 [51]. This also gives rise to the popular theoretical construct of the folding funnel, which depicts the energy landscape during folding as a multitude of energy gradients the protein must pass to reach the lowest possible energy state[52] (and thus its native conformation). While this concept first proposed by Ken Dill in 1989 is helpful in visualizing the core principles of protein folding, its two-dimensional nature and coarse scope is inadequate to describe the multiple folding pathways a protein can take. A way to improve on this concept has already been proposed in the 90s [53]. Shifting the perspective from a single pathway the protein can take to a three-dimensional landscape resembling a rugged crater, displaying valleys and crevices that can describe local energy minima, or even intermediates respectively[23,54].

However, these schemes were developed to describe protein folding that has been investigated mainly *in vitro*. The environment for proteins to fold *in vivo* have been described as vastly different, possibly influencing the folding landscape a great deal[55,56,57].

In the last two decades, protein design *in silico* became a powerful method to create macromolecules of pre-mediated function. The rise of new methods to predict, design and reliably produce de-novo protein structures utilizing machine learning techniques has opened many new avenues for scientists to create novel protein structures for specific functions[58]. We are currently at the precipice of a new age of protein design and engineering, with a great potential for progress[59,60,61,62]. But with computing resources currently being the limiting factor, access to this technology is restricted to those with the computing infrastructure to support it[50].

1.3. Fragments and their significance in protein evolution

Investigating the origin of modern protein structures is a tricky endeavor, and we can only rely on deduction of plausible scenarios of what happened based on evidence that we can gather today. Thankfully, continuous progress in related sciences (like for example cryo-EM) made it more accessible than ever to evaluate great amounts of data on protein structures, their sequences, and the wealth of additional information. With the rise in computing power in the last 20 years, statistical analysis of the entirety of the known protein universe is well within our capabilities. Already early on, the

recurrence of certain sequence or structural features in proteins from diverse backgrounds led to the concept of shared ancestry being a major component in the evolution of proteins[63,49]. Through mechanisms like duplication, permutation, fusion, and general mutation, a set of existing primordial stretches of proteins could have given rise to most of the modern protein folding space[64,65,66,67].

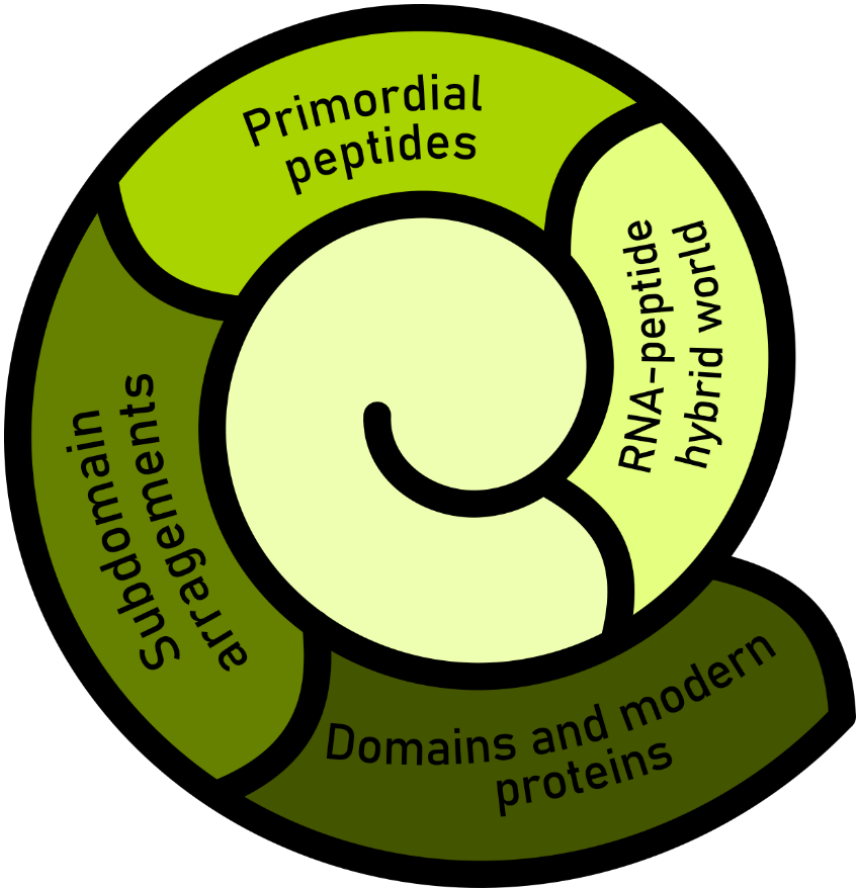


Figure 3: Step-by-step process of the evolution of complex protein structures from smaller building blocks. Starting from a theoretical RNA-peptide hybrid world where structure could have been governed by interaction with specific RNA, the emergence of the first self-folding peptides, their diversification via sub-domain arrangements and finally their assembly into modern multi-domain proteins. This schematic represents just one interpretation of how protein structures could have evolved, based on information of each successive step.

Due to the nature of how we think protein evolution happened – the combination of smaller, pre-existing building blocks – a better understanding of the underlying processes is of great interest (Figure 3). Classification and grouping of protein structures according to their structural and sequence properties and subsequent inferrance of evolutionary relationships is a well-established approach and can be used to investigate some key aspects in the evolutionary hierarchy of protein structures. Protein structural networks used in the annotation of proteins like SCOP[68,69], CATH[70]

or ECOD[71,72] can offer a wealth of information on protein structures and provide an easy to use but powerful tool for gathering information on a given protein. Additionally, these extensive databases can be used in training machine-learning approaches, incorporating evolutionary information into the resulting output[73,74,75,76].

The mentioned databases, however, have one thing in common, the smallest unit they consider is that of the protein domain. To include the smaller, sub-domain building blocks or fragments, newer methods in the classification of sequence and structure relationships can be implemented to generate a more detailed evolutionary hierarchy of these structural elements. Introduction of sequence analysis methods such as hidden Markov models (HMMs) in combination with the data already accessible via the established structural classification made it possible to detect remote homology that had previously been inaccessible for analysis[77,78,79]. Taking these methods into account led to a paradigm shift in how protein evolution must be considered. Detection of remote homologies using sensitive sequence analysis methods, shows that elements previously thought to have stemmed from convergent evolution most likely share a common, homologous origin[78,80].

Several databases exist that try to classify these new-found relationships within folding space and represent some of the first steps in systematically mapping this subdomain regime of protein evolution[81,82,80]. Identification and examination of subdomain fragments produces a clear picture: In line with the previously proposed mechanisms of early protein evolution, re-use of an extensive fraction of sequence space is prevalent across the entirety of the protein fold space. Implementation of this wealth of data will help us solve the question of the evolutionary origin of some of the most ubiquitously occurring protein folds in nature[83]. A categorization of these ancient building blocks is also useful for protein design. If we follow the idea that in general, these fragments proliferated because of their desirable properties of foldability, they could also introduce a suitable scaffold to design on. Proteins re-using already established building blocks in such a way could also benefit from the native-like properties, such as a certain flexibility. This approach could open avenues for the design of proteins with future models potentially profiting from this intrinsic flexibility.

One of these databases trying to catalogue the evolutionary and structural relationship of protein fragments is the Fuzzle database[80]. Fuzzle is based on the SCOPe database and utilizes HMMs created from sequence information of about 60% of all PDB entries

and the SCOP domain classification[73]. This sequence information was then used to infer structural relationships. By comparing similarities in structure using both the absolute deviation in structure (RMSD) of sequences aligned in the HMM analysis and superimposing the results using TM-align[84], it was possible to identify sub-domain sized regions in this comprehensive dataset (Figure 4). The results of this analysis were subsequently classified by similarity and clustered into different units, termed *fragments*.

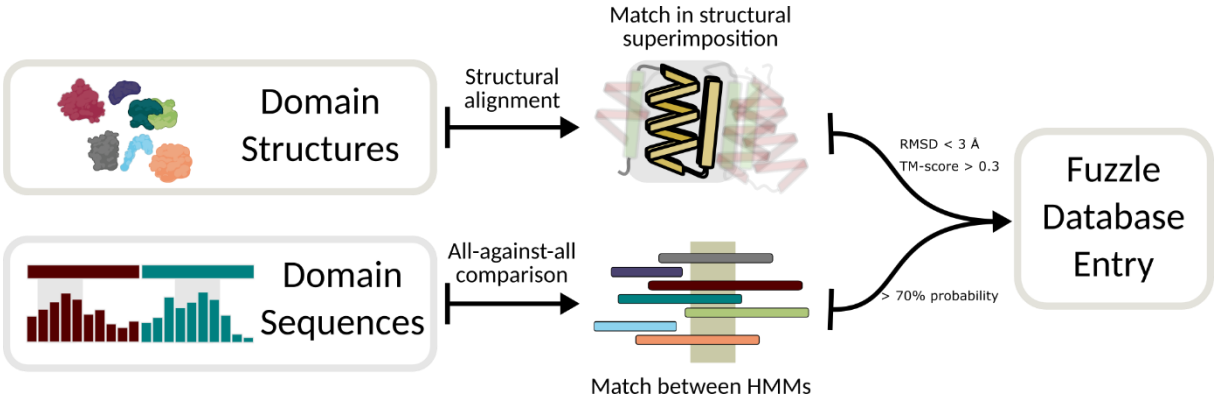


Figure 4: General construction of the Fuzzle database. Structure and sequence information from the original protein dataset is compared via structural alignment and profile-profile comparison of sequences. The results of these two steps are then concatenated in a singular database and filtered with specific cut-offs. Shown here are the standard cut-offs.

1.4. Fragments in the context of entire proteins

It has been shown that modern-day proteins arose from a combination of different evolutionary mechanisms like duplication, permutation, fusion of sequences and mutations[85,86]. This implies that before these mechanisms could create the fold space we observe today, there had to already be some smaller elements, perhaps governed by these same rules[87]. Indeed, the implementation of sequence[79,88] and structure analysis sensitive enough to detect these remnants of evolutionary ancestry show stretches of sequence in modern proteins that are shared between folds that were previously not considered to be evolutionarily connected[42,82,89,90]. This could also be one of the reasons why we observe a conservation of transition-states in the folding of small proteins from homologous families[91].

The relationship of sequence space and the connections we observe at a structural level for subdomain-sized fragments can now be addressed by utilizing these new methods. A visualization of this network as clusters of relationships in combination with the already existing domain classification (for example using SCOPe as underlying information) highlights the high interconnectivity of these fragments across the domain space[92]. Taking the Fuzzle database as an example, the fragments that are detected span a considerable portion of the protein fold universe. Using the standard cut-offs in this database (70% HHM probability and a TM-score of over 0.3) and using the resulting dataset to create a network representation offers a suitable way of depicting the interconnectivity at play. Of the roughly 28000 domains classified in the Fuzzle dataset, over 8 million pairwise hits can be found, meaning that there are a significant number of hits spanning more than one domain[80]. This large amount of data can however be clustered based on the similarity of its individual hits. For example, a fragment consisting of only a few secondary structure elements can be found to have been re-used in many different folds, while still hinting to a homologous origin (Figure 5). Identification of fragments that can be considered unique can then be mapped onto the known fold space (which in the case of Fuzzle is SCOP), resulting in a network of interconnected fragments within different folds.

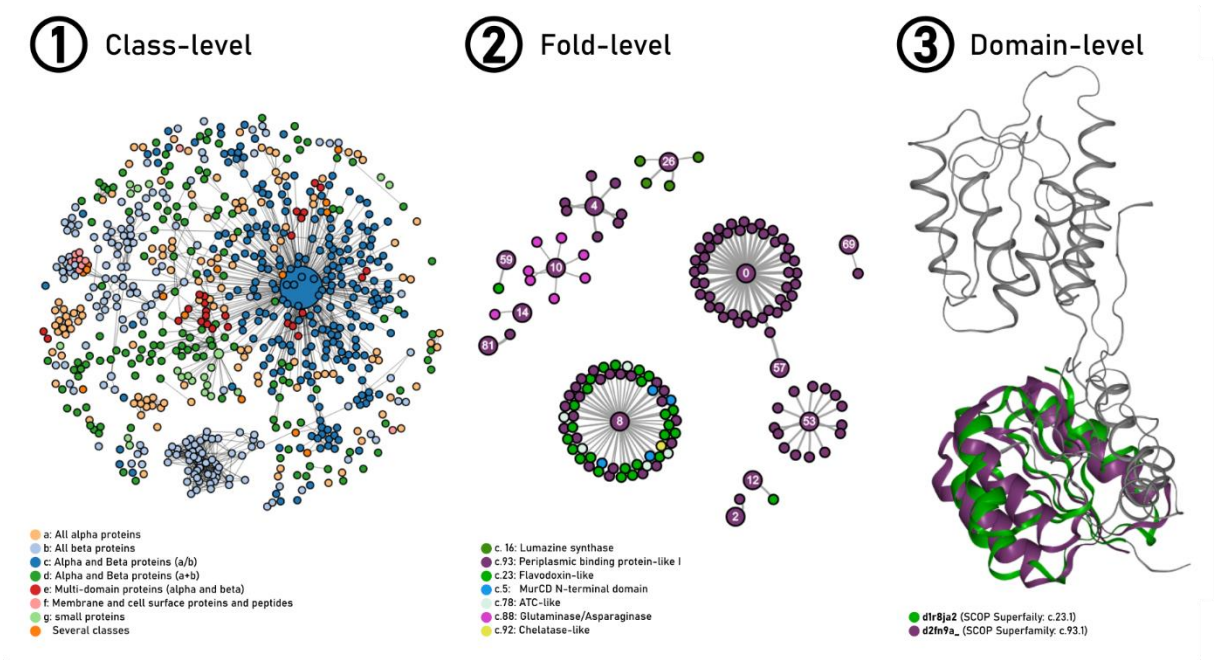


Figure 5: Different structural levels that can be explored hierarchically with help of the Fuzzle database. (1) On the class-level, hits within Fuzzle are clustered in a connected network of folds. Connections between classes rise from a connected fragment found in Fuzzle. (2) On the fold-level, information on connection between folds is accessible. Shown here is the connection of a fragment taken from the ribose-binding protein domain (SCOP-ID: d2fn9a_). (3) On the domain-level, information of the fragment is broken down to a structural level. The structural alignment of the fragment is shown in context of the parental proteins.

This clustered network can not only help show the possible evolutionary connections of a given fragment within the context of multiple different protein folds, but also be used to find suitable fragments for the insertion or grafting of elements in a different context. This technique has been successfully applied in the construction of protein chimeras, and in the future could be a way to even transfer functions from one protein to another[90].

1.5. Evolution of a protein – The case of periplasmic binding proteins

One of the folds that has been in the focus of evolutionary and protein engineering investigations for a long time is the periplasmic binding protein-like fold. The group of periplasmic binding proteins (PBPs) consists of a range of bacterial solute binding proteins that are involved in the binding and transport of various ligands classified into different types (see 1.6 for the more detailed classification). Their role is not only the transport of solutes into the cell, but also the associated signal transduction makes them play an important part in the ability of an organism survive[93] (Figure 6).

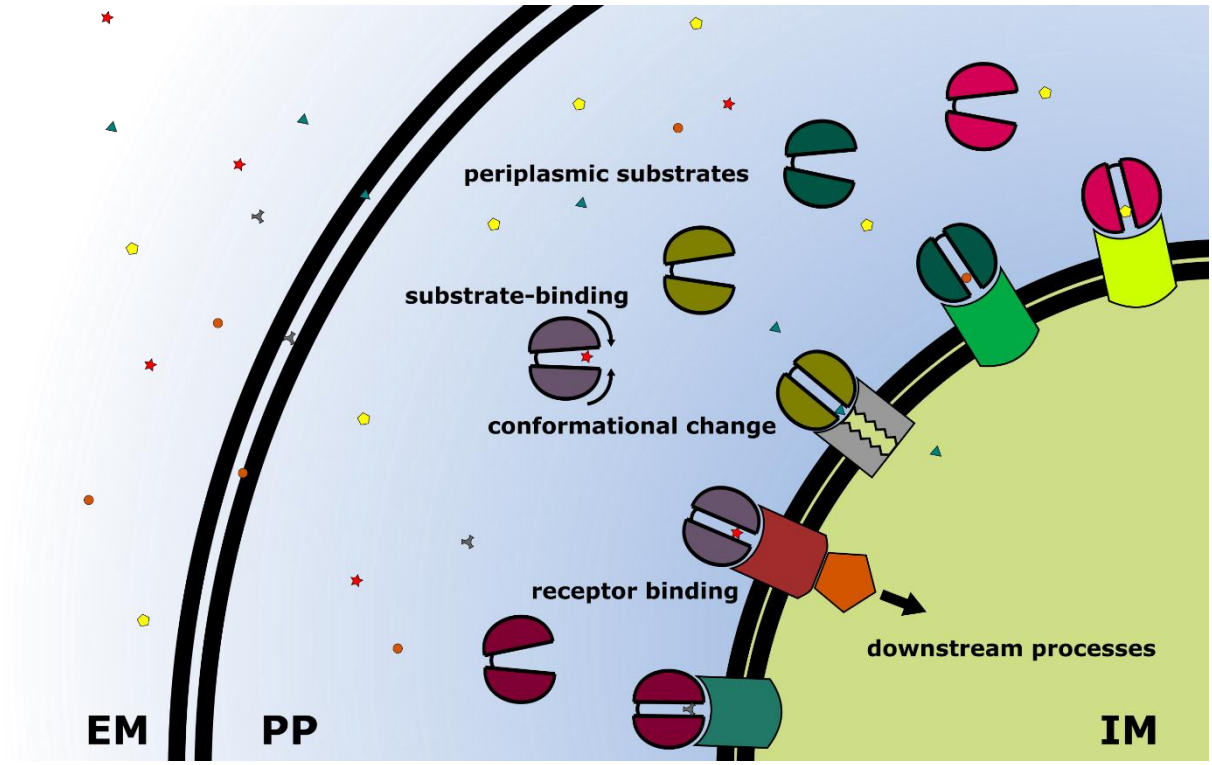


Figure 6: Overview of the mode of operation of periplasmic-binding proteins. Solutes (different small symbols) are transported from the extracellular milieu (EM) via pores into the periplasm (PP). Periplasmic proteins in the PP then recognize each specific solute and undergo their characteristic conformational change upon binding. This enables them to bind to their cognate receptors, leading to various downstream processes.

Their versatility and importance in many central cell functions is further reflected in the variety of ligands PBPs can recognize and bind, such as mono- and oligosaccharides, amino acids, short peptides, minerals such as sulphates, phosphates or vitamins[94]. Their mode of action is often described as a result of their architecture resembling a "Venus-flytrap"-like shape[95]. PBPs consist of two lobes connected via a short hinge region, with the ligand binding site being formed in between those two domains. Unbound PBPs generally exist in an equilibrium of an open and closed conformation, with the two states being defined by the change in angle of the two lobes on both sides of the ligand binding site. Under physiological conditions, this equilibrium is almost entirely on the side of the open conformation, allowing for the specific ligand to bind. Upon binding the equilibrium however is strongly shifted to the closed conformation, effectively trapping the ligand between the two lobes. In this closed conformation, PBPs are able to interact with the corresponding membrane-bound receptors located at the surface of the cell, resulting in the respective outcome, such as the active transport of the solute, or downstream signal.

This central role in the metabolism of both gram-positive and gram-negative bacteria is one of the possible reasons why this fold is found ubiquitously within this group of organisms, and why this class of proteins adapted to binding such a wide variety of solutes[96,97,98,99]. The unique bi-lobal architecture that is classified as the PBP-like fold (named after the functional class of bacterial proteins) can also be found in numerous eukaryotic proteins[100]. The PBP-like fold as a module appears as a binding motif in proteins, like seen in the crystal structure of the glutamate receptor GluR2, and based on sequence similarity is proposed to also exist in other hormone- and neurotransmitter-binding domains[101]. This relationship of the binding modules of eukaryotic binding proteins and the PBP-like proteins highlights the pervasiveness of this particular fold being found in nature.

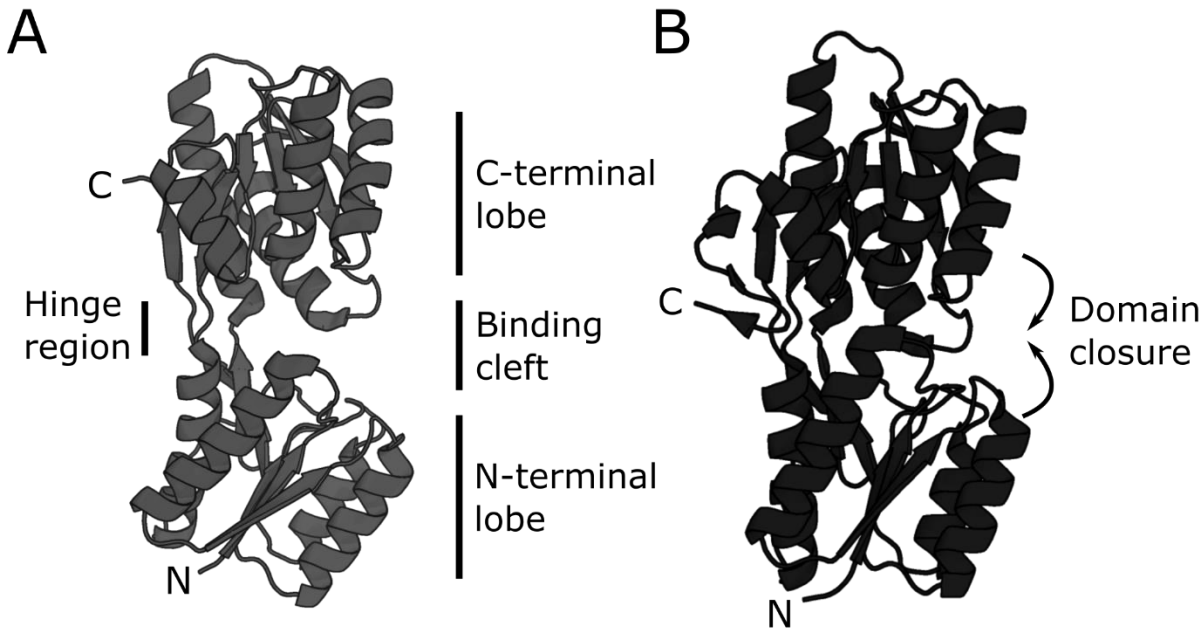


Figure 7: Schematic structural makeup of a periplasmic binding protein. Characteristic bi-lobal structure consisting of the opposing N- and C-terminal domains, with the binding cleft in-between. **(A)** Open structure of the PBP. Opening and closing of the protein is facilitated by the hinge region at the opposite end of the binding cleft. (**B**) Closed structure of the same PBP, with the closure of the cleft around the ligand. (Structures used are ribose bound form [PDB-ID: 2FN8] and unliganded form [PDB-ID: 2FN9] of Ribose Binding Protein of *T.maritima*.)

The versatility of functions can partially be explained by the unique topology of the PBP-like fold[102]. As previously mentioned, the binding site of the ligand is located in the cleft between the two globular lobes, with a transition of the PBP from an open to closed conformation upon binding. The two opposing lobes each consist of a parallel five-stranded β-sheet, flanked by an alternating pattern of five α-helices, with the two lobes being connected via a short linker region. In the open conformation, the angle between the two lobes allows for solvent to access the interface between them. If the

specific ligand is recognized, the subsequent binding event induces a conformational change in the hinge region, resulting in a large shift in the orientation of the two lobes towards the closed conformation. This “closing” of the interface not only allows for correct orientation of the interacting residues from each side of the lobe, but also helps in excluding solvent from the binding site[103]. This mechanism of action offers several advantages, from securing the ligand in an environment free from the influence of surrounding solvent, as well as the thus tightly controllable interface allowing for single-point mutations to fine-tune ligand interactions, and the conformational change on the surface of the entire protein enabling recognition by the respective downstream receptors.

This specific mode of binding makes PBPs an attractive target for biotechnological applications[94]. Their ability to accurately recognize specific ligands and bind them with high affinity can be used to engineer complex reporter systems for a variety of small solutes. Additionally, the extensive conformational change can be utilized to enable insertion of functional groups within the protein, creating for example a photometrically detectable readout upon binding. A range of biosensors has already been created this way, detecting a variety of different solutes using the natural affinity of the corresponding PBPs in combination with an additionally engineered readout module[104,105,106]. Since then, several improvements to specificity and sensor sensitivity have been made[107]. Further improvements to engineering approaches utilizing the PBP-like fold also enable the direct readout of carbohydrate concentrations in-cell[108], or the active monitoring of glucose[109]. Other approaches for molecular engineering utilize the ability of PBPs to interface with transporters located in the cell wall of bacteria to actively transport unnatural amino acids into the cell[110], or use the ability of PBPs in quorum sensing to monitor solute concentration in a bacterial culture via FRET-readout[111].

However, these techniques use only the natural affinity of already existing PBPs, modulating readout by introducing additional residues on the surface of the protein. Extensive efforts have been conducted to redesign the ligand specificity of this class of protein, particularly to function as novel biosensors for the application in a biologically relevant context, for example the detection of solutes like neurotransmitters *in vivo*[112,113]. The immense diversity of this protein fold highlights the potential of technical applications, and explains the intensive effort put into engineering PBPs to suit a specific function. The ubiquitous occurrence in nature and the resulting great

number of possible templates is already a good starting point. In combination with the already existing functions and the precise mode of binding while still offering enough malleability to do redesign makes it a good target for continued investigation.

1.6. Structural classification of PBPs

Generally, PBPs share low sequence similarity with each other, making classification of this class a complex issue by itself[114]. PBPs are classified into three types, based on differences in their topological arrangements and sequence similarities. Based on the number of crossovers within the hinge region, PBPs are generally categorized into either being Type I, II or III. These three types correspond to clusters found in sequence analyses[115,116] and are used in the main structural classification databases. However, there are also other studies on the classification of PBPs, for example based solely on structure and binding specificity, or the combination of structural and sequence information, which lead to the definition of multiple different structural clusters[117,118,119,120]. Every one of these classifications are in agreement with the general distinction into Type I – III. There do not appear to be any incongruities in the overlap of the different classifications with those identified as Type I-III. The periplasmic binding proteins are part of the class of solute binding proteins, which in turn are part of the larger group of the ATP-binding cassette (ABC) protein superfamily. Similar to the ubiquitous nature of PBPs, members of the ABC superfamily have been found in all phylogenetic branches, with the different functions of the cassette often being found as paralogues across the genome[121,122,123,124]. ABC proteins perform a variety of functions, ranging from the previously mentioned importing of periplasmic solutes in bacteria, or ATP-driven import of substrates, DNA-repair and translation regulation or transmembrane transport of hormones, lipids, peptides or other secondary messengers in eukaryotes[100]. Analysis of different sequence databases indicates that the genes for ABC proteins most likely stem from multiple gene duplication and fusion events, causing the significant diversity we see in this gene family today[125,97,96].

1.7. Dismantling the protein – the evolution of periplasmic binding proteins

Already shortly after the first structures of PBPs (the solute binding part of the ABC transporter cassette) were solved the hypothesis was put forward that this fold might have evolved via duplication of a smaller single domain protein, an idea first put forward

by Louie *et al.*[126] and further explored in the classification by Fukami-Kobayashi *et al.*[115]. However, since PBPs are considered to be a relatively old fold, sequence identity has likely drifted too far to infer any direct homologous relationship with other folds. At least when utilizing relatively straight-forward sequence analysis methods such as BLAST, no homology with other folds can be detected.

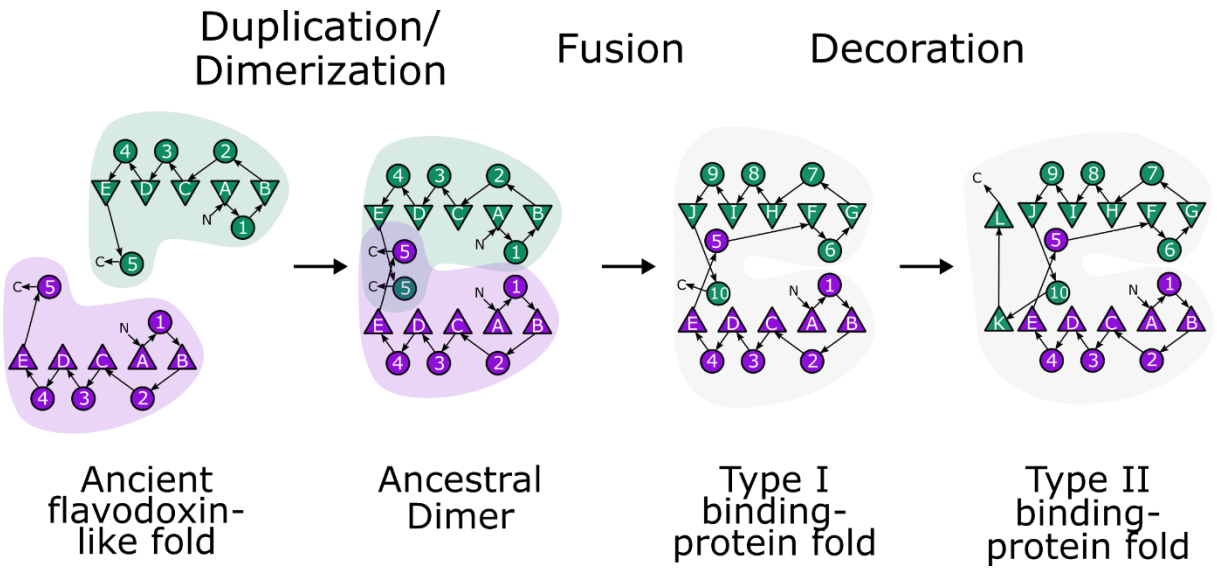


Figure 8: Proposed evolutionary trajectory of modern PBP-type I proteins and the derived constructs used in this study. *Proposed steps that reconstruct the evolution of modern periplasmic-binding-protein (PBP) folds from an ancestral protein adapting the flavodoxin-like fold (adapted from Fukami-Kobayashi, 1999). A duplication and dimerization along with swaps in secondary structure led to the formation of an ancestral dimer. Subsequent fusion of the genes then led to the emergence of an ancestral PBP-like fold and further changes of secondary structure to that of the modern PBP-like type I fold. (Figure adapted from* [127]*)*

There is some evidence that the PBPs of type II originated from a domain dislocation step from those of type I. This is corroborated by the intersections of these types observed in the more detailed classification methods mentioned in chapter 1.6. The differences in sequence and in the arrangement of secondary structure elements led to the hypothesis that a tandem domain swap was the responsible event leading to the divergence of the two classes[115]. Connecting the emergence of specific structural insertions, accretions, and deletions with the proposed evolutionary age of the PBPs further enabled inferring of directionality of evolution. With both PBP types being found in bacteria and archaea, the divergent event must be older than the separation of those two clades if an origin from a homologous ancestor is presumed. The number of in-depth studies on the general evolution and these structural peculiarities makes the PBP-like fold an interesting candidate for the investigation of early evolutionary events.

1.8. Thinking smaller – Origin and elements of the PBP fold

As previously mentioned, it has been proposed that the modern architecture of PBPs arose from the duplication of a single $(\alpha\beta)_5$-protein. Based on structural similarities, candidates have been proposed to originate from the family of CheY-like proteins. The distinctive mode of binding could have been facilitated by the subsequent exchange of an α-helix from the originating domains into the opposing one, creating the proto-hinge region that led to the venus-flytrap like behavior. However, this event would have had to happen early in the evolution of PBPs. The considerable divergence of possibly related sequences due to its age makes it extremely difficult to substantiate this claim. So far, no analysis led to a definite conclusion, but studies on the domain arrangement[128] and studies on the evolutionary relationship of similar folds[90] offer reasons to further investigate this case.

1.9. Fragments in the evolution of PBPs

Considering the evolutionary history of the PBPs, originating from a duplication event of a primordial CheY-like protein, we can also investigate the relationship between sub-domain fragments found in this fold. Since analysis methods allow for comparison of fragments found in all folds, inferring possible evolutionary relationships between folds should be possible as well. This approach has previously been used to identify not only the relationship in other protein folds like the $(\beta/\alpha)_8$-barrel or the TIM-barrel fold and the flavodoxin-like fold[129], or the HemD-like fold and the flavodoxin-like fold[90]. Both studies used the identification and subsequent characterization of these fragments as a way to investigate the evolutionary relationship of the shared sub-domain unit. Using these fragments that are believed to have a common evolutionary origin is a way to overcome the insufficient evidence of homology when using the entire protein in sequence comparisons.

One of the proteins where this approach has been used is the TIM-barrel fold. This protein consisting of eight βα-elements is formed by the symmetrical assembly of the eight β-strands into a central barrel, with the associated α-helices forming the outer surface of the barrel. It has also been proposed that this protein evolved via a duplication of a precursor equivalent to one half of the barrel[130]. This relationship is also found when applying deep sequence searches utilizing hidden Markov models. When using these methods, a clear relationship between TIM-barrels and proteins of

the flavodoxin-like fold can be detected[131], hinting at a common evolutionary relationship (Figure 9). These connections can further be investigated with a hybrid-approach using the detection of common fragments through available databases such as Fuzzle in combination with experimental data.

The design and investigation of chimeric proteins – meaning proteins in which a portion of the sequence is exchanged by another evolutionarily related one – can be one such tool to experimentally investigate the properties of fragments thought to be related. Proteins containing pieces from different protein folds have already been created and characterized using fragments from the flavodoxin-like fold with the TIM-barrel fold[132,133], the HemD-like fold ([134,135]) and also the PBP-like fold (PDB 4QWV). Not only does elucidation of the structure confirm correct folding in most of these chimeras, but it is also possible to retain original binding capability, verifying the similarity of the inserted fragment in a possibly physiologically relevant context. Since all these chimeras rely on an "illegitimate recombination"[132] of a fragment thought to stem from an evolutionary precursor, and thus being from a smaller protein one can also turn that approach around: Taking a modern version of a protein fold and try to isolate either the halves (presuming duplication), or the fragments (presuming recombination). A stabilization of sub-domain elements within a fold has for example already been observed for the HemD-like half[90]. One of the challenges of this approach is the significant divergence of the sequences from their progenitors. Due to evolutionary pressure on stabilizing interactions being lifted the isolation of the individual fragments or subdomains can be difficult[136].

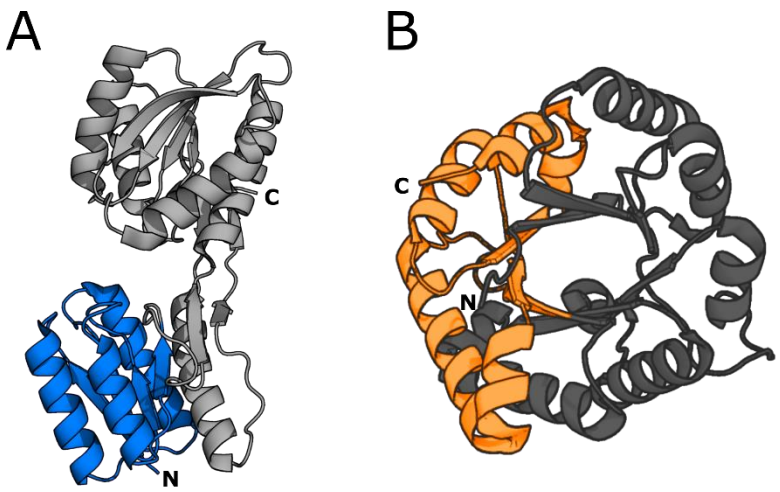


Figure 9: Exemplary fragments found in a PBP-like fold and a TIM-barrel-like fold. (**A**) Fragment found in the Ribose-binding protein of *T.maritima*.(PDB-ID: 2FN9) (**B**) Fragment found in the triose phosphate isomerase from *G.Gallus (*PDB-ID: 1TIM*)*. While not directly related, both fragments share similarities with the original protein they were found in, as well as share connections to the flavodoxin-like fold.

To summarize, we now not only have the tools to detect evolutionary relationships between fragments of proteins of different folds, but also have methods at our disposal to investigate these relationships in more detail. If we now apply the same approach to the PBP-like fold, we can get a more comprehensive picture of the evolution of some elements within this fold. This study aims to recapture the process of fragment propagation, incorporation into a possible flavodoxin-like ancestor and the following duplication leading to the modern PBP-like fold I. Additionally, the investigation of a particular fragment detected in the PBP-like fold I can provide more information on why some fragments are so widely detected in many different folds. To this end, the modern ribose-binding protein of the thermophilic organism *Thermotoga maritima* and the main fragment found within that protein have been investigated in the context of its possible evolutionary trajectory. While it is almost impossible to ever provide any definite proof of the evolutionary history of PBPs as millions of years passed since the events described here transpired, this study provides useful indicators to expand the already fascinating progression of this fold.

This work is a comprehensive study of an evolutionarily related fragment, which has been taken from within the context of its parental protein and analyzed regarding its evolutionary significance. This could give important insights in the principles of early protein folding. One of the most interesting avenues could be the investigation whether these fragments could pose the illusive fold-on units important for early-stage protein folding. As at least the fragment investigated in this work seems to fold, adopt structure, and is resistant to extensive mutation of its sequence this approach could work for the investigation of other fragments as well. With the existence of sub-domain databases, a systematic study of different fragments could further our understanding of these elements, with possible applications for targeted protein design down the lane.

2. Synopsis

2.1. Dismantling of an established PBP

With the rise of more powerful computational methods and the wealth of structural information we have at our disposal, we can now try to retrace the pathway of evolution from fragments to a modern PBP-like protein. As with most studies in the field of protein evolution however, there is the caveat that the evidence we gather only leads to an inferred conclusion. Since we cannot directly access structures or sequences already existing millions of years back, we must use indirect methods, accessible via investigating the nature we *can* observe today.

Consider for example the modern PBP-like fold as a starting point. When we look at its structure, one first notices its general shape. It can be roughly distinguished into two, almost equal portions, connected by a smaller cross-over region. If we then look closer at the sequence similarity of these two structurally similar lobes, their sequence similarity does not support direct inference of a common ancestry[137]. This holds true even when comparing a variety of individual lobes of the known PBP-like sequences from different sources among each other. One possible explanation to this low similarity could be that the fold originated from two individual proteins, which at some point fused (Figure 10) [115].

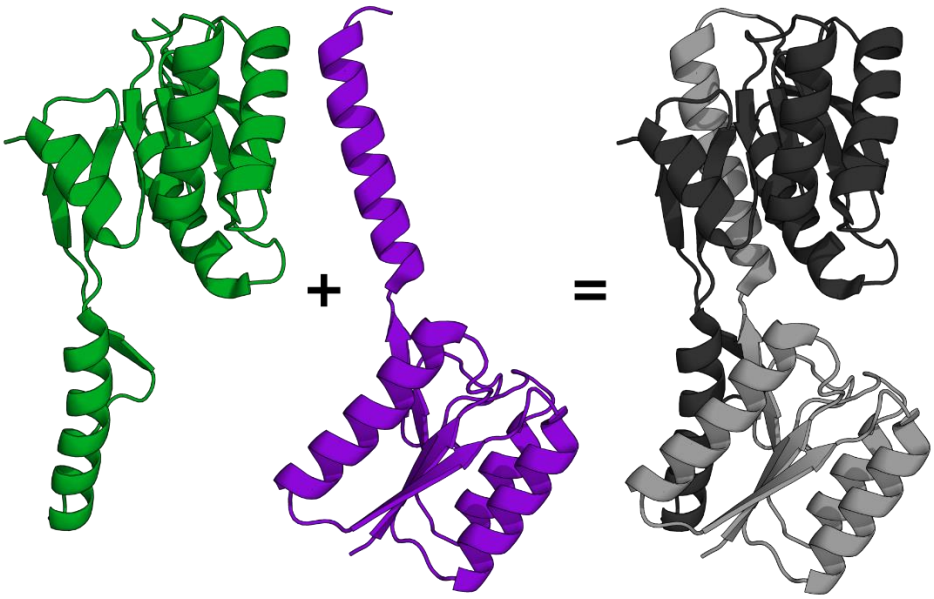


Figure 10: Showcase of the two principal components of a modern PBP. *To the left are the two halves of the ribose-binding protein (PDB: 2FN9) identified by using HHpred sequence analysis. To the right is the complete structure. Worth noting is the symmetry of the two lobes, as well as their structural similarity with the flavodoxin-like fold.*

An alternative explanation for the evolution of this fold from two or more unrelated elements is also possible. Since evolution is guided by efficiency, often navigating the simplest possible path to a given solution and factoring in the existing structural and (to an extend) sequence similarities make this a less likely explanation. One should keep in mind when trying to explain an origin of this fold from two (or more) individual proteins: To reach the bi-lobal structure always necessitates the fusion and adaption of at least two components. An alternative explanation comes to mind if we take a closer look at the sequences. By not only using multiple sequence alignments but combining that approach with the building of profiles we can make more detailed comparisons of sequences. This in turn enables us to improve our predictions of sequence similarities based on the probabilities of any given amino acid in a profile[79]. If implemented, this approach enables us to see even remote homologies of the PBP-like fold. While we cannot directly detect a similarity of the entire sequence of the halves with each other, we find a common link: Proteins of the flavodoxin-like fold are detected when looking for remote homologies of the single lobes. This supports the idea that PBPs and flavodoxin-like proteins might share a common evolutionary origin (see chapter 1.7). These results also provide additional credibility to the proposal of the duplication of an ancestral flavodoxin as the origin of the PBP-like fold[138], similar to what has already been described as a common origin for other folds[90].

To investigate this process in detail, a modern Ribose-binding protein was analyzed in depth, with the hypothesis in mind: "*If modern PBPs originate from a duplication event, could it be possible to revert this process, and end up with a singleton resembling a flavodoxin-like protein?"*

This is something we explored in Paper I, with the disassembly of the Ribose-binding protein from *T. maritima* into its two constituent lobes and investigating their structure and function both in isolation as well as in combination.

In Paper II, an alternative approach to this disassembly was implemented, using permuted variants of the two lobes, identifying their structural differences to the proposed model, and opening new implications for the flexibility of the PBP-like fold.

2.1.1. Paper I – The two lobes

A method that has been employed to investigate the role of sequence duplication in the generation of new protein topologies is the dissection of modern folds into parts[139,140,141]. Applying the same methodology to investigate the properties of stability and binding of the individual lobes of the modern ribose binding protein (RBP) from *T.maritima* can similarly help us understand its proposed evolutionary history. Coupled with modern sequence analysis, this structural investigation can inform us not only on the duplication event itself, but also the process of the folds divergence into its multitude of functions after the evolutionary and functional decoupling of the individual lobes[142,143,144,145].

Since previous studies on the duplication event have been based on either structural[114,118,146,117] or sequence[138] analysis, a more hands-on investigation can add further insight. To this end, RBP of *T.maritima* was used as a model system. In addition to the protein being more accessible for analysis due to its thermophilic nature, a truncated version of this protein was reported indicating that it might be more amenable to structural manipulation[103]. To investigate this duplication event, several constructs based on the sequence analysis of the modern RBP were designed and characterized biochemically.

To generate the different constructs, the information of the multiple sequence alignment, the corresponding HHpred[77,147] profiles and structural features were taken into consideration (Figure 11).

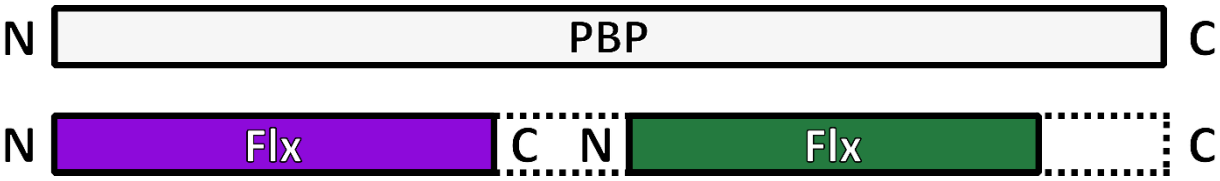


Figure 11: **Schematic of the alignment of the two flavodoxin-associated profiles found within the sequence of the RBP.**

The resulting analysis of the constructs clearly showed that not only it is possible to stabilize both individual lobes of RBP, but that their behavior is even like that of the full-length protein. Furthermore, experiments with the combination of the different constructs indicate that it is possible for the individual lobes to regain their binding function when in presence of each other, while this is not the case when tested in

isolation. Even when the two corresponding halves were purified independently, mixing them reconstituted binding function to a degree similar to that of the parental protein. The successful crystallization and determination of the structure of the two parts in presence of each other shows a dimer of the two lobes very closely resembling the structure of the full-length RBP. These results are in agreement with the observed ribose binding, indicating that the protein reconstitutes its native conformation to a point that re-enables function. Upon binding of ribose, the thermostability of the heterodimer of the N- and C-terminal halves increases significantly, almost exactly mirroring the behavior of the intact, full-length RBP. In addition, previous constructs that only expressed insolubly can be brought to solubilize when co-expressed with the corresponding half via the formation of a dimer.

The fact that we can create stable proteins based on the halves of a modern PBP-like protein is a good indicator for a duplication being the reason. While the sequence of the individual lobes has diverged to a point that direct inference of homology has become impossible, the structural similarity and sequence analysis by profiles still supports this hypothesis.

2.1.2. Paper II – The permutations

Another approach implemented to study the two lobes of RBP has been the introduction of permutations within the sequence of the halves. Core principle of this study is the swap of the α-helix 4 and 9, respectively in-between the two lobes of the modern PBP-like fold. If we think of the two lobes stemming from a single entity that has been duplicated somewhere on the evolutionary timeline of what is now a two-lobed PBP, the helices should – in theory – also stem from the same structural element of the progenitor. This means that these helices already possess a sequence optimized to interact with the equivalent surface of each lobe. Arguably, this enables us to isolate another analogue of the ancient single-lobe configuration of the progenitor.

In this new approach, a permutation between each swapped α-helix and its corresponding β-sheet on the same side of the binding cleft structurally isolates each lobe – albeit in a non-linear fashion to its sequence. The two constructs generated in this manner thus represent the structural entities of the N- and C-terminal lobe of RBP (Figure 12).

The points of permutation were manually selected after analysis of structural and sequence data. A comparison of the full-length protein with proteins of the flavodoxin-like fold that are thought to be modern day ancestors of the single-domain protein at the origin of PBPs was used to inform the point of duplication. After determination of most likely duplication sites, the structures were cut at these positions, and the resulting fragments connected via short, computationally designed loops of 3 to 4 residues. With this permutation in place, it was possible to isolate and individually produce the two lobes of RBP that at least structurally represent the units of the progenitor prior to duplication. Subsequent analysis of their biochemical features showed that both halves form stable, well folded proteins in isolation (Figure 12). An interesting result of this study was that the lobes show a strong intrinsic propensity to form stable, well-defined dimers. For example, the N-terminal construct (*RBP-CPN*) displays a concentration-dependent equilibrium of monomeric and dimeric states in solution. The C-terminal construct (*RBP-CPC*) shows similar behavior, with a shift of monomeric population to dimer, albeit at higher total concentrations than its correspondent half.

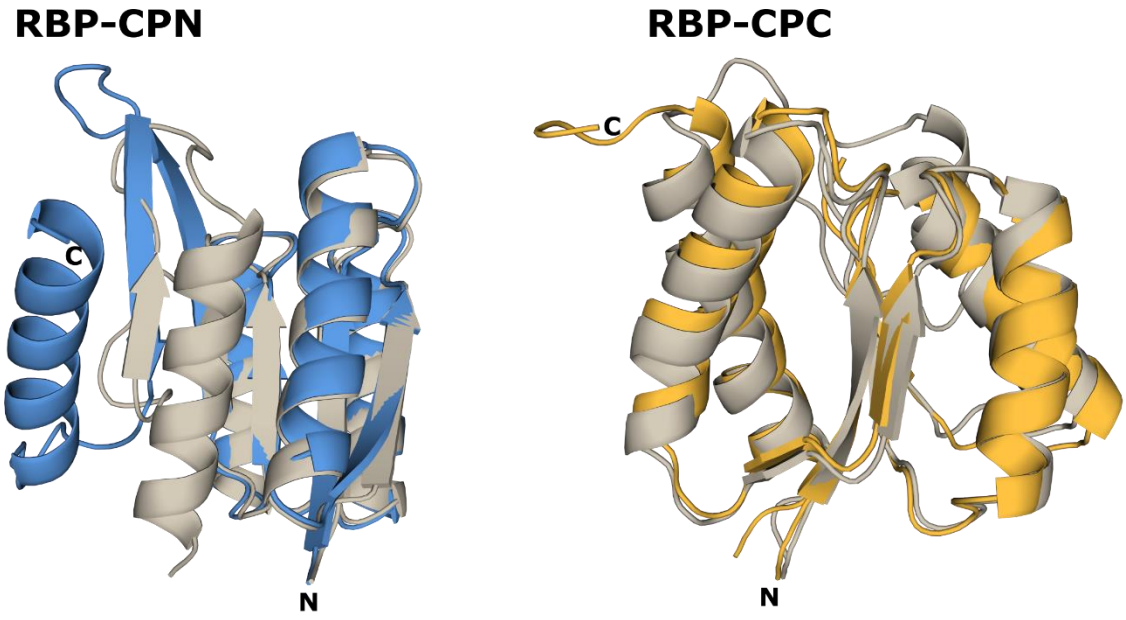


Figure 12: **Crystal structures of single-chains of RBP-CPN (left, blue) and RBP-CPC (right, yellow).** Comparison with the model structures generated with Rosetta (in grey) shows an almost identical core structure, with some differences observable in the C-terminal region of both the N- and the C-terminal permutated structures.

In both halves however, the formation of the dimer increases the thermodynamic stability of the constructs, as shown by the significant increase in the transition temperature in DSC measurements. Determination of the atomic structures of the two proteins confirms the formation of the dimer for at least RBP-CPN. However, both RBP-CPN and RBP-CPN display previously unobserved configurations of secondary structural elements. Dimer formation within RBP-CPN is facilitated by a novel swap of the C-terminal βα-elements of each monomer, resulting in an antiparallel elongation of the central β-sheet. This configuration has not been observed before in modern proteins of the PBP-like or flavodoxin-like folds and gives some insights into how flexibility of structural rearrangements can create platforms for the evolution of new protein topologies.

2.2. Fragments and their role in the PBP-like fold

While the proposed duplication event of an ancestral progenitor most likely played the major role in the generation of the functionality of modern PBPs, it is not the full evolutionary story we can tell. Utilizing the powerful analysis tools which have been developed in recent years, we can infer even more connections in the evolution of modern proteins. In this case, using the Fuzzle database it was possible to identify a fragment within the N-terminal part of RBP, spanning its first 88 residues. Looking at the fragment in a network of folds different from the originating PBP-like fold II, we can detect this fragment as part of a cluster of considerable connectivity. One of these connections we observe within this fragment is to proteins of the flavodoxin-like fold. This is another indication that the duplication of an ancient flavodoxin-like protein is at the origin of the PBP-like fold. However, within the cluster some previously unknown connections to other folds can also be detected. This would imply that the origin of this fragment might even predate the origin of the flavodoxin-like fold. To investigate the connections within the Fuzzle database in more detail, we took a closer look at the relationships of the fragment between different folds, and the functions carried in these contexts.

The next step was to investigate the biochemical and structural properties of the fragment. Isolating the residues equivalent to the fragment found in RBP and removing it from its structural context led to several constructs with interesting properties. The

existence of these elements in isolation opens interesting implications for the existence of these fragments, as well as their role in the evolution of proteins in general.

2.2.1. Paper III – Using Fuzzle as a tool for identifying fragments

The Fuzzle database allows to search for such common fragments in different contexts. The database combines the structural information within the already established SCOP database and their sequences. SCOP offers the possibility to use information on the homology of proteins on a domain level, classified in a hierarchal manner. The relationships in SCOP are leveled as

- **families** (clear evidence of shared evolutionary origin),
- **superfamily** (mostly same structure, probable evolutionary ancestry),
- **fold** (grouping via shared structure, not necessarily related),
- **class** (classification by secondary structure content only)

These terms are generally used to infer evolutionary relationships (at family and super-family level) or structural similarities (at fold and class level) of protein structures[73]. If we however combine this structural information with sequence analysis that is sensitive enough to detect even remote homologies – in this case Hidden-Markov-Model (HMMs) based sequence comparisons, we can access an additional layer of information[148]. Using an all-against-all comparison of the HMM-profiles generated with the entire SCOP dataset, it is possible to identify matching regions. Depending on the cut-offs used in this analysis, the likelihood of these matching regions sharing a common evolutionary ancestry is high. The main advantage of this approach is its capability of finding matching regions on relatively small stretches of different sequences without major loss of prediction accuracy.

Combining these profiles with the data already contained within SCOP, it is possible to assign each of these unique matches to certain structural elements. As a last step, a sequence superposition of the profiles within the structures in SCOP at a certain similarity cut-off offers another indicator for the proposed evolutionary relationship.

Combining all three information levels, the sequence profiles, the evolutionary hierarchy from SCOP, and the structural comparison resulted in the distinct set of fragments we observe in the Fuzzle database[80].

One additional aspect of classifying these subdomain-sized fragments as common evolutionary units is their hypothetical interchangeability within existing protein structures. Several chimeric proteins have been successfully created using this approach. One of the main goals of these studies has been the conservation of function within the fragments and bringing them into a new structural context. Since the previous version of Fuzzle (1.0) did not include ligand information, an update to include molecule interactions was developed, allowing for systematic searches of ligands in the dataset. The research paper not only shows an update of the underlying SCOP-dataset to a newer version increasing the total amount of hits within the database, but also highlights the possibility of finding relevant functionalities within any given fragment.

To showcase this new capability, the N-terminal fragment of RBP was analyzed in detail regarding its connection to other proteins and possible ligand interactions. Not only is the fragment displaying a higher than usual connectivity in the entire network – meaning that it is found within proteins of many different folds – but also a wide variety of different ligand binding. The enhanced analysis showed that the fragment can be found in 121 unique protein domains from a total of 15 superfamilies and 9 folds. Through the inclusion of ligand binding information in the 2.0 release of Fuzzle, it was possible to identify 21 unique domains sharing a connection with the fragment in the network and their accompanying ligands.

One of the main goals of Fuzzle is to provide a database for identifying remote evolutionary relationships in a sub-domain regime. However, the additional functionality the ligand analysis provides can also be used to inform more functionally inspired endeavors. The creation of functional protein chimeras by active site transfer via switching the entire fragment could be a possible application. Additionally, the information of functionality and evolutionary connections could be used to inform evolutionary analyses in the future.

2.2.2. Paper IV – Isolation of Fragments from RBP

While the previously mentioned events of duplication and rearrangement of elements within a given protein structure are the main drivers of diversification of protein structures today, evolution of structures at the beginning of the protein universe were governed by different rules. It has been proposed that a specific subset of suitable substructures – called fragments – posed the primordial starting point of protein structure. While the specific identity and propagation of these fragments is a matter of great debate and generally hard to follow through the course of evolution, several ideas and concepts have been developed to classify this region of sequence and structure space. Modern sequence analysis coupled with the plethora of structural information made it possible to attribute recurring elements within different protein folds to a possible shared evolutionary origin. Utilizing this method, we identified one of those fragments within the modern PBP-like fold II of RBP. This fragment of 88 residues consists of the first two αβ-elements of the full-length RBP.

To understand more about the possible role this fragment played in the structure of the protein today, and consequently gain more insight into its relevance in the evolution of this fold, we looked at the fragment in isolation i.e. away from its structural framework. To this end, variants consisting of only the fragment were designed and analyzed (Figure 12). One of the constructs is the fragment directly taken from the sequence of RBP. In this case, this were the first 88 residues of RBP from *T. maritima*. Since the ancestry of this fragment cannot only be found within the PBP-like fold it originates from but other folds as well (see also 4.2.1.), the hypothesis arises that its identity might predate the divergence of these folds.

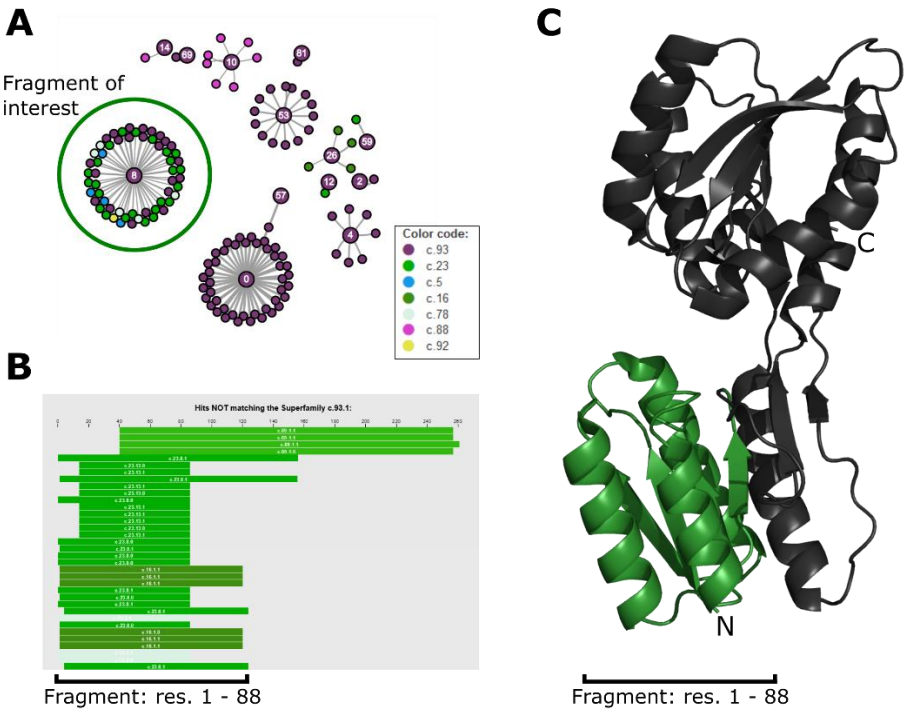


Figure 13: **Analysis of the fragment within RBP.** (**A**) Cluster representation of the different fragments identified using Fuzzle. In this case, Fragment No. 8 corresponds to the fragment shared with the flavodoxin-like fold. (**B**) HHpred results taken from Fuzzle, with the sequence of the fragment corresponding to the first 88 residues of the whole RBP sequence. (**C**) Cartoon representation of the structure of RBP (PDB: 2FN9) with the fragment highlighted in green.

We can show that isolation of the fragment is not only possible but appears to yield a relatively stable protein in solution, with spectroscopic data displaying behavior similar to that of the parental full-length protein. Light scattering analysis also indicates the protein is monomeric at low concentrations, with a concentration-dependent shift to a possible dimer configuration at higher concentrations.

To get an idea of how this fragment might be adaptable to change, we used the consensus sequence as an approximation of a more ancient-like sequence. While ancestral sequence reconstruction could also have offered a way to analyze the behavior of this fragment, the extensive sequence analysis to create a suitable phylogenetic tree was out of scope for this study. However, since the profiles generated during the HMM-search provide a multiple sequence alignment with stringent cut-offs, they also provide suitable consensus sequences for the fragment. Utilizing this approach, we created two fragment constructs. The consensus fragment (cFragment) from the sequence considering all folds found in the HMM-analysis, and the consensus

fragment with the consensus only built from sequences not of the same fold-class of PBPs (cFragment$_{noPBP}$).

Both consensus fragments can be obtained from the soluble cell fraction upon expression in *E.coli*, however only cFragment shows proper secondary-structure formation. The results of light scattering analysis also indicate that while cFragment can be found as a homodimer in solution over a wide range of concentrations, cFragment$_{noPBP}$ does not have any defined peak indicating a lack of globular structure.

While there are still many open questions about this fragment, the results clearly show that it is not only possible to isolate the parts from the originating PBP, but also that it forms a stable protein unit. The behavior of the fragment also closely resembles that of the parental protein, indicating that it might form a comparable structure as well. Additionally, changing the identity of this fragment by introducing mutations via a consensus approach introduced some interesting behavioral changes in the fragments. For the cFragment the secondary structure content still seems to be comparable to that of the original fragment. However, it appears to form a stable dimer in solution, while still retaining its secondary structure. It is impossible to tell whether this is a random byproduct introduced by the changes in sequence, but it would be interesting to investigate a possible relationship of whether this propensity to form multimers might be an intrinsic property of fragments. Understanding the mechanisms of this could help understand the origins of bigger proteins by accretion. If other fragments share this property, it could partially explain the propagation of fragments in fold space.

2.3. Further exploration of these concepts and their application in protein design

While we can now reliably identify segments of proteins that are very likely to share an evolutionary connection, we still lack the fundamental understanding of why we can find these recurring elements. There exist several theories as to why these elements – even when not directly incorporating any function – might have been successfully reused within evolution. One such concept is that protein diversification simply started with a very limited subset of structural archetypes, and through duplication, accretion, diversification, mutation, and deletion in addition to the hundreds of millions of years of

evolution led to the complexity we observe nowadays. Since this also implies that all this stems from only a handful of amino acid sequences, it would follow that we are still able to identify and describe these relationships today. As to why some of these fragments do not appear to carry any function, a possible explanation that is often brought up in this context is the existence of an RNA-based world[149]. In this scenario, the essential catalytic functions stem from an interacting ribozyme partner, with the structural backbone being provided by a non-functional peptide. It is possible that the fragments we observe today might be remnants of those architectural features, rather than functional ones.

This however does not exclude the possibility of these fragments carrying any inherent function. The process of fragment propagation in the early evolution of protein structures could have happened both before and after the first catalytic functions carried out by the proteins themselves. Functional residues could also have evolved within any fragment already incorporated in the bigger context of a larger protein as well, resulting in the various ligand binding observed in fragments found in Fuzzle 2.0.

Combining the structural uniqueness of the fragments, their proposed role in early protein evolution and their functions in the context of their modern counterparts could greatly help inform protein design in the future. Understanding what makes these fragments especially attractive to be kept in a protein, even without any functional benefit could help us understand what is important for overall protein fitness. As these fragments might display interesting properties regarding scaffolding or general stability of a protein, investigating their behavior could also result in principles that may be used in protein engineering or design.

2.3.1. Paper V – Connecting the fragments

An important aspect of investigating these evolutionary mechanisms is their applicability to protein design. Understanding how the fragments work in the context of the entire folding space of proteins can help us develop a mix-and-match approach to function without the need for extensive *de-novo* design. Consequently, understanding the underlying mechanisms of these evolutionary events also helps us in classifying the modern fold space. To use this information of not only the origin of fragments, but also their mechanisms of propagation can lead to new insights of how folds arose. This

can also be applied to learn common principles in protein folding, for example whether fragments can generally represent independent folding units.

To review the current state of the field of evolutionary informed protein design, this work sets out to summarize the different aspects on the example of a single fold. The fold that was chosen for this purpose is the TIM barrel, a symmetric barrel-shaped fold. Similar to what has been shown with the PBP-like proteins, the TIM-barrel proteins are thought to stem from a duplication event not unlike that proposed for the PBPs.

Starting with a comprehensive overview of evolutionary mechanisms, we highlight the possible ways a protein must undergo, starting from small evolutionary units, their diversification, and how we can apply knowledge of these processes for protein design applications. By also highlighting other attempts to classify the subdomain fold space, our own work with Fuzzle is set into perspective in an emerging field of protein structure classification. Using this knowledge of the subdomain parts, we tried to retrace the evolutionary history of the TIM barrel, focusing on its connection with other folds of similar α/β proteins. Putting it into context of other publications, a way to identify several possible steps in the evolution of TIM barrels is proposed. To highlight the interconnectivity of evolutionary processes, their implications in protein structure and stability as well as manipulation of the TIM barrel fold, an overview of the current state of literature on both evolution and design of the fold is given. Additionally, application of these past conclusions and the *de-novo* design of a protein with a TIM barrel architecture is discussed.

To consolidate the aspects needed to be considered when trying to design or engineer an existing protein, the four main aspects of evolution, folding, stability and function are set in the context of this ubiquitous protein fold.

3. Author contributions

Paper I: Retracing the Evolution of a Modern Periplasmic Binding Protein

F.M., S.R.R., B.H. designed the research, F.M., S.R.R. purified the constructs, F.M. collected and analyzed CD, IF, and SEC-MALS data, S.R.R. performed DSC and ITC experiments, F.M., S.R.R. solved X-ray structure. All authors wrote the manuscript.

Paper II: Structures of Permuted Halves of a Modern Ribose Binding Protein

F.M. performed the design of the permuted constructs of both RBP-lobes, the stability analysis *in silico*, the cloning, and protein purification. The analysis via CD, IF and MALS was also done by **F.M.**, with help from S.S. and S.R.R. Setup and optimization of the protein crystallization was handled by **F.M.**. S.S. performed crystal mounting, data collection, structure determination and refinement. S.R.R. conducted the DSC-experiments, as well as the thermodynamic analysis. The study design was done by B.H. and **F.M.**. B.H. further provided supervision and funding. All authors contributed to the writing of the manuscript.

Paper III: Fuzzle 2.0: Ligand Binding in Natural Protein Building Blocks

N.F., **F.M.**, F.L., S.S., and B.H. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Paper IV: Isolation of Subdomain-sized Elements in a Modern Periplasmic Binding Protein

F.M. and B.H. designed the research, **F.M.** and T.K. purified the different constructs, collected, and analyzed CD, IF, and SEC-MALS data, and performed the DSC experiments. **F.M.** and B.H. wrote the manuscript.

Paper V: Evolution, Folding and Design of TIM Barrels and Related Proteins

This review was conceptualized and designed by all participating authors. Primarily **F.M.** contributed to the section about categorization of the different sub-domain units and the associated databases, in addition to providing feedback for the rest of the manuscript.

4. Paper I

Michel, F., Romero-Romero, S., Höcker, B.
Retracing the Evolution of a Modern Periplasmic Binding Protein.
Protein Science, 2023, 32(11)

Received: 10 July 2023 | Revised: 20 September 2023 | Accepted: 22 September 2023

DOI: 10.1002/pro.4793

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

Retracing the evolution of a modern periplasmic binding protein

Florian Michel | **Sergio Romero-Romero** | **Birte Höcker**

Department of Biochemistry, University of Bayreuth, Bayreuth, Germany

Correspondence
Birte Höcker. Department of Biochemistry, University of Bayreuth, Bayreuth 95447, Germany.
Email: birte.hoecker@uni-bayreuth.de

Funding information
Alexander von Humboldt-Stiftung; European Research Council; Volkswagen Foundation

Abstract

Investigating the evolution of structural features in modern multidomain proteins helps to understand their immense diversity and functional versatility. The class of periplasmic binding proteins (PBPs) offers an opportunity to interrogate one of the main processes driving diversification: the duplication and fusion of protein sequences to generate new architectures. The symmetry of their two-lobed topology, their mechanism of binding, and the organization of their operon structure led to the hypothesis that PBPs arose through a duplication and fusion event of a single common ancestor. To investigate this claim, we set out to reverse the evolutionary process and recreate the structural equivalent of a single-lobed progenitor using ribose-binding protein (RBP) as our model. We found that this modern PBP can be deconstructed into its lobes, producing two proteins that represent possible progenitor halves. The isolated halves of RBP are well folded and monomeric proteins, albeit with a lower thermostability, and do not retain the original binding function. However, the two entities readily form a heterodimer *in vitro* and *in-cell*. The x-ray structure of the heterodimer closely resembles the parental protein. Moreover, the binding function is fully regained upon formation of the heterodimer with a ligand affinity similar to that observed in the modern RBP. This highlights how a duplication event could have given rise to a stable and functional PBP-like fold and provides insights into how more complex functional structures can evolve from simpler molecular components.

KEYWORDS

flavodoxin-like fold, gene duplication, protein evolution, ribose binding protein, solute binding protein

1 | INTRODUCTION

The detection of chemicals in the environment, their molecular recognition and transport into the cell as well as the resulting downstream signaling is an integral part of life in any cell. As one of the central classes of proteins responsible for this function in prokaryotes, the periplasmic binding proteins (PBPs) serve as an important element in these complex response networks (Matilla et al., 2021). These bilobal proteins are involved in the transport of a wide variety of substrates, and are generally considered to belong to an

Florian Michel and Sergio Romero-Romero contributed equally to the work.

Reviewing Editor: Aitziber L. Cortajarena

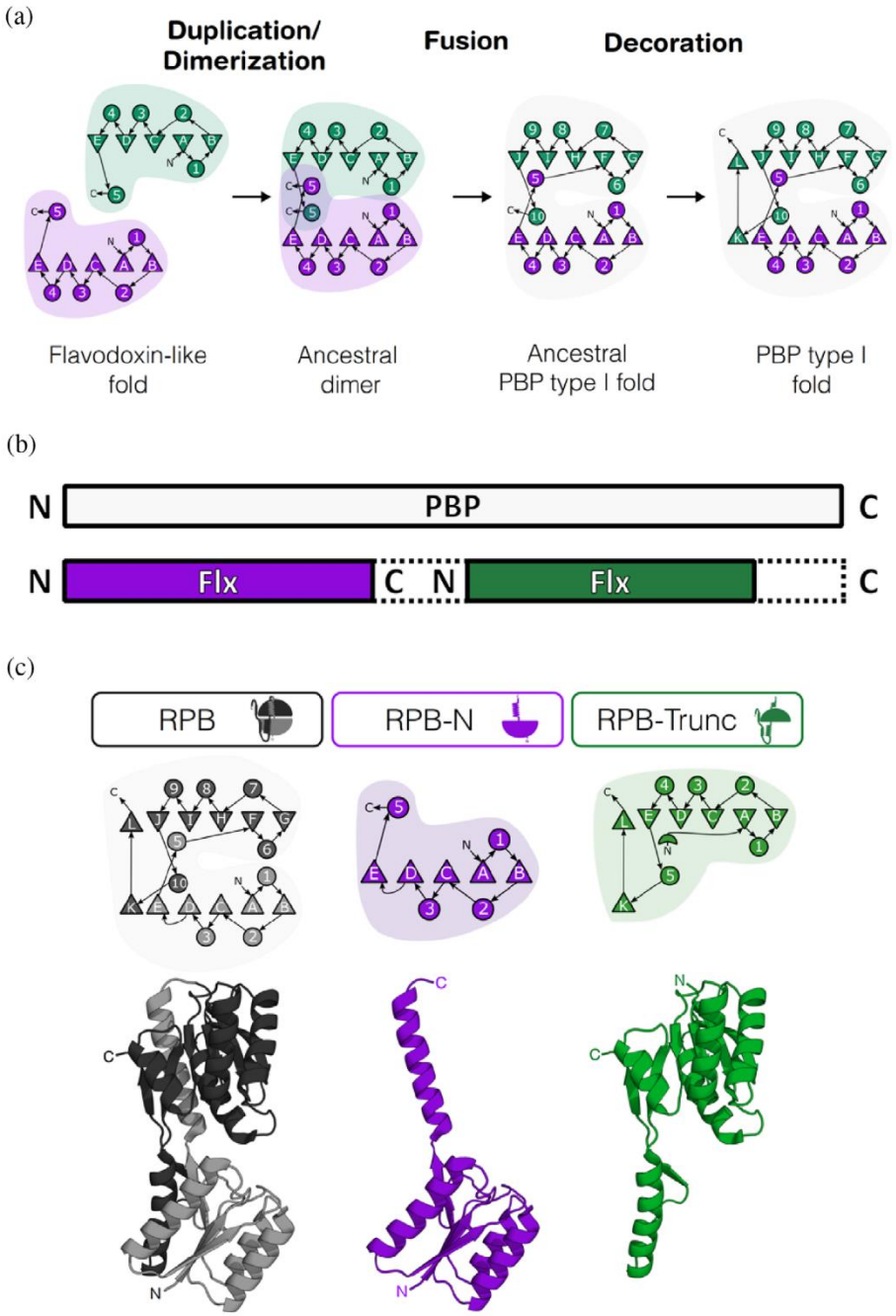


FIGURE 1 Proposed evolutionary trajectory of modern PBP-type I proteins and the derived constructs used in this study. (a) Proposed steps that reconstruct the evolution of modern periplasmic-binding-protein (PBP) folds from an ancestral protein adapting the flavodoxin-like fold (adapted from Fukami-Kobayashi et al., 1999). A duplication and dimerization along with swaps in secondary structure led to the formation of an ancestral dimer. Subsequent fusion of the genes then led to the emergence of an ancestral PBP-like fold and further changes of secondary structure to that of the modern PBP-like type I fold. (b) Schematic representation of the profile-profile alignments for a representative full-length PBP with the flavodoxin-like-fold (Flx). (c) First-generation constructs RBP (black), RBP-N (violet), and RBP-Trunc (green) were analyzed in this work to recreate the PBP-halves.

ancient protein fold (Clifton & Jackson, 2016; Felder et al., 1999).

The PBP architecture consists of two opposing lobes, with each lobe being built of a central, five-stranded parallel β-sheet with five α-helices flanking its sides. The two lobes are connected via a hinge region, with the complexity and number of crossovers dependent on the class of PBP. This architecture also gives rise to the most common mechanism in which PBPs recognize and bind their ligands (Berntsson et al., 2010; Chandravanshi et al., 2021; Scheepers et al., 2016). This distinct mode of binding that a majority of PBPs follow is a "venus flytrap-like" mechanism and considered one of the hallmark features of this protein class (Felder et al., 1999). While in the unbound state, PBPs are in an "open" form with a space created by the two lobes accessible to surrounding solutes. Recognition and binding of the ligand facilitates interaction between the two lobes, leading to the eponymous hinge-bending motion which results in the "closed" conformation with the cleft now being tightly shut around the ligand, excluding the solvent upon binding (Berntsson et al., 2010; Felder et al., 1999). This common binding mechanism is reflected in PBPs that bind similar molecules with very

different selectivities and affinities at the same binding site (Kröger, Shanmugaratnam, Ferruz, et al., 2021). For these reasons, PBPs have been used in several engineering and design approaches, especially creating highly sensitive biosensors and molecular switches (Dwyer & Hellinga, 2004; Jeffery, 2011; Medintz & Deschamps, 2006; Steffen et al., 2016), and designing new binding properties (Banda-Vázquez et al., 2018; Kröger, Shanmugaratnam, Scheib, et al., 2021; Scheib et al., 2014).

Despite diversity in the sequences of different PBPs, a shared common ancestry has been proposed a while ago (Fukami-Kobayashi et al., 1999; Louie, 1993). Their structural features, similarities in binding mechanism, and shared operon structure—with the PBP being on the same operon as the associated signaling proteins downstream—have long led to the theory that PBPs arose via gene duplication of a progenitor protein and subsequent diversification. However, it is unclear in which order these events might have occurred (Fukami-Kobayashi et al., 1999). It has been previously suggested that this common ancestor could have been a CheY-like protein adopting a flavodoxin-like fold. Formation of an ancestral dimer in combination with a gene duplication and fusion event might have led to the typical bilobal structure of the modern PBP (Figure 1a), an event that has already been investigated for the evolution of other protein folds (Alvarez-Carreño et al., 2022; Farías-Rico et al., 2014; Toledo-Patiño et al., 2019).

Although the sequences of modern PBPs have diversified from their evolutionary ancestors, the topology is predominantly conserved. There are mainly two classes of PBP, with a slight difference in the order of secondary structural elements. It is thought that the second class descends from already evolved class I PBPs even though sequence similarity is not high between the two folds (Fukami-Kobayashi et al., 1999). They are in fact classified as independent folds of either PBP-like I or PBP-like II in SCOP (Chandonia et al., 2019), as being of the same topology level as flavodoxins (for type I) and an independent homology group (for type II) in ECOD (Cheng et al., 2015), and as two different superfamilies in CATH (Sillitoe et al., 2021). The application of modern bioinformatic resources has opened up new opportunities to revisit some of these concepts of evolutionary relationships, partially through emergence of tools to more efficiently probe sequence space also in the sub-domain regime of proteins (Alva et al., 2015; Farías-Rico et al., 2014; Ferruz et al., 2020; Nepomnyachiy et al., 2017).

In this work we combine the approach of a sequence profile-profile comparison analysis using Hidden Markov Models (HMMs) with a structural comparison of the two lobes of the PBP-like fold type I. Based on this analysis, the emergence of the PBP-like fold via the duplication of a flavodoxin-like ancestor can be revisited. To further substantiate the claim, we biophysically and structurally characterized truncated constructs of the ribose-binding protein (RBP) from *Thermotoga maritima* that correspond to the proposed duplicated progenitor halves. We found that it is generally possible to obtain stable and well folded monomeric proteins expressing only the individual lobes of full-length RBP. The two independent halves appear to readily form a heterodimer, while also reconstituting the ribose-binding ability of the parental protein, with affinities in the same order of magnitude. These results suggest a plausible path for the evolution of modern PBPs and increase our understanding of the evolution of complex and multidomain proteins from smaller molecular components.

2 | RESULTS AND DISCUSSION

2.1 | Disassembling a modern RBP into likely progenitor halves

The proposed mechanism of a duplication event being responsible for the architecture of PBPs mostly relies on analysis of either the available structures of modern PBPs (Louie, 1993; Berntsson et al., 2010), or comparison of the sequences of PBP-like and flavodoxin-like proteins (Fukami-Kobayashi et al., 1999). We wanted to investigate whether the duplication of the flavodoxin-like progenitor is not only theoretically feasible, but also practically. To retrace the evolution of a PBP, we characterized constructs based on the halves of an RBP (Figure 1 and Table S1). This not only allows to probe the plausibility of this mechanism in general, but also offers an opportunity to investigate the individual impact of each subdomain-part on the stability and function of modern PBPs.

We chose the RBP of *T. maritima* for this purpose. Not only does the thermophilic nature of this protein offer a robust system, but also a previously reported expression of a 21 kDa truncated version (Cuneo et al., 2008) made this an excellent candidate for a model system. To generate an overview of possible intersections, a multiple sequence alignment with RBP as input was generated with HHpred (Figure S1). The results show not only the alignment of other full-length PBPs on the query sequence but also an alignment of the individual lobes. The lobes align with a clear cut being observable between residues 30–155 and 156–310 of the RBP (numbering consistent with Uniprot entry Q9X053). To compare this with the alignment of the proposed progenitor

flavodoxin-like proteins, the same alignment was generated within the *Fuzzle* database (Ferruz et al., 2021), which automatically excludes sequences of the same fold. It shows that flavodoxin-like proteins align with both the corresponding N- and C-terminal halves of the PBP sequence (Ferruz et al., 2021). While alignment of flavodoxin-like proteins with RBP seems to heavily favor hits on the N-terminal half, some hits are also found with the C-terminal half. A reason why less hits might be observed on the C-terminal half of this modern RBP could be a result of the duplication and a subsequent decoupling of the sequences of the two halves, resulting in increased divergence from the progenitor flavodoxin-like protein, and thereby making it harder to identify.

While the existence of the earlier reported truncated RBP variant could be an artifact of the expression in *Escherichia coli* (Cuneo et al., 2008), it is also possible to be a natural occurrence. A shortened version of a solute-binding protein with a proposed biological function has been reported previously (Bae et al., 2018). Although it is unclear why these single-lobed proteins might exist, the truncated RBP could also carry biological significance. Thus, we chose to use the truncated protein that is roughly the equivalent of the single-lobed half as a base for the constructs used in this study.

For the first generation of constructs we took to the lab, the sequence of the full-length RBP was disassembled into the corresponding halves (Figure 1 and Table S1). The site of dissection was determined by structural alignment of RBP in absence of ribose (PDB ID: 2FN9) to the top-scoring flavodoxin-like proteins in the HHpred analysis, resulting in the constructs RBP-N (amino acid 30–153 of RBP) and RBP-C (amino acid 157–291) that contain a sequence identity and similarity to each other of 16.8% and 25.6%, respectively. These constructs were expressed and characterized using biochemical and biophysical methods.

2.2 | RBP halves are well folded

Upon overexpression of the RBP halves in *E. coli* the protein RBP-N was found in the soluble fraction of the cell extract while RBP-C was located in inclusion bodies. Since full-length RBP also features a C-terminal decoration common to modern PBPs which does not correspond to any elements in the canonical flavodoxin-like architecture, the additional elements (two β-strands that facilitate another cross-over between the two lobes and extend the central β-sheet of the two halves) had been removed in RBP-C. This removal might be the reason why in contrast to RBP-N, which expressed solubly, could be purified to homogeneity, and remained stable at concentrations above 15 mg mL^{-1}, RBP-C only expressed insolubly. We therefore decided to continue the investigation with the truncated construct RBP-Trunc (residues 142–310) instead (Figure 1c), which is related to the RBP-C half and expressed solubly with similar stability to the N-terminal construct RBP-N.

Both RBP-N and RBP-Trunc display far-UV CD spectra with the signature α-helix minima at 208 and 222 nm and moderated by the signal of the β-sheet at 218 nm, both characteristic for α/β-proteins (Figure 2a) and comparable with the native full-length RBP. Comparison of the intrinsic fluorescence (IF) also corroborates this (Figure S2A), indicating that the constructs are well folded since the intensity maximum suggests that the aromatic residues are buried from the solvent. In addition, DSC endotherms show cooperative thermal-unfolding transitions with melting temperatures and enthalpy values close to full-length RBP (Table 1, Table S3, and Figure 3), confirming the characteristics of well-folded proteins (see next section for further details).

Further analysis with SEC-MALS (Figure 2b) confirmed the monomeric state of RPB and RBP-N. However, RBP-Trunc is in an equilibrium of mostly monomeric species and homodimers, with higher oligomers also being present (Table S2). These results indicate that the RBP halves are well folded proteins and express mainly as monomeric systems, similar to those observed in another PBP, HisJ (Chu et al., 2013). To follow up on this, we continued to study their properties in the presence of each other.

2.3 | RBP halves form a heterodimer whose structure is identical to full-length RBP

Since one of the steps proposed in the evolution of the modern PBP architecture involves an ancient dimer, we investigated whether the obtained constructs had the ability to reconstitute the full-length RBP fold. For this, the individually purified RBP-N and RBP-Trunc were mixed in an equimolar ratio and then analyzed. The far-UV CD spectra (Figure 2a) show a significant change of the signal to the individual constructs, with the signal of the mixed RBP-N/RBP-Trunc resembling that of the full-length RBP. A similar behavior can be observed in the IF spectra (Figure S2A), where the original characteristics of the full-length protein are reconstituted when mixed *in vitro*, hinting at the formation of an RBP-N/RBP-Trunc heterodimer. Complex formation is supported by SEC-MALS analysis where only one well-defined peak is displayed corresponding to the mass of the RBN-N/RBP-Trunc dimer of heterodimers (Figure 2b and Table S2).

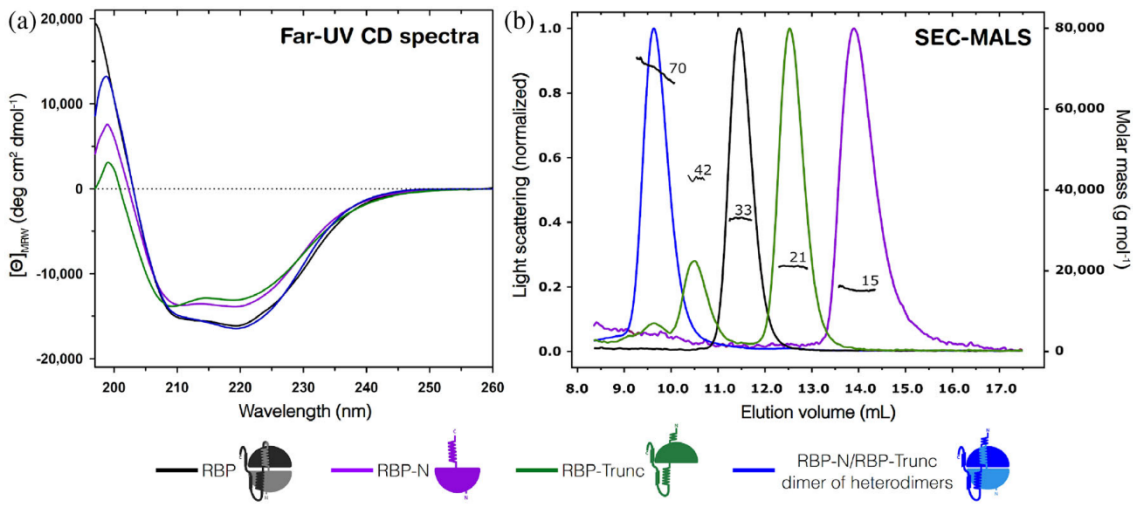


FIGURE 2 Biophysical characterization of the first-generation constructs. (a) Far-UV CD spectra at 20°C collected in 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8. (b) SEC-MALS experiments performed in 10 mM sodium phosphate, 50 mM sodium chloride, 0.02% sodium azide, and pH 7.8. Numbers indicate the molecular weight determined after data analysis. Values derived from the experiments are reported in Table S2. In both panels, the color code is RBP (black), RBP-N (violet), RBP-Trunc (green), and the RBP-N/RBP-Trunc dimer of heterodimers (blue).

Additionally, DSC analysis of the proteins supports the formation of a heterodimer that resembles the parental protein. All endotherms show clear single and cooperative transitions, as has been observed for other PBPs such as maltose-, arabinose-, and histidine- binding proteins (Fukada et al., 1983; Ganesh et al., 1997; Kreimer et al., 2000). However, RBP and its halves showed irreversible thermal unfolding possibly due to their thermophilic nature, contrary to most PBPs which exhibit reversible transitions (Aggarwal et al., 2011; Fukada et al., 1983; Ganesh et al., 1997; Kreimer et al., 2000; Prajapati et al., 2007; Vergara et al., 2020). While full-length RBP has a T_m of 106.9°C similar to the one previously reported for the construct (Cuneo et al., 2008), RBP-N and RBP-Trunc show lower thermostability with a T_m of 76.6 and 73.3°C, respectively (Figure 3a, Table 1, and Table S3). The results show that the halves have native-like properties, that interdomain interactions are important in RBP and that these provide relevant stabilization, in the same way as has been described for other multidomain proteins (Brandts et al., 1989; Careaga et al., 1995; Kantaev et al., 2018; Liu et al., 2019; Vergara et al., 2023; Vogel et al., 2004; Wenk et al., 1998). This decrease in thermostability of the individual constructs is compensated by the formation of the RBP-N/RBP-Trunc heterodimer, whose T_m is shifted by more than 20–99.7°C, more closely resembling that of RBP.

The same tendency is observed when comparing the changes in ΔH of the individual and mixed constructs (Table S3), with a considerable increase of 240 kcal mol^{-1} in the unfolding enthalpy, which is significantly higher than only the sum of the individual halves (115 kcal mol^{-1}). These differences indicate that more accessible surface area is exposed upon unfolding, which is most likely due to the formation of an extensive interface and interdomain interactions important for protein stability and function as present in RBP, confirming the interaction between RBP-N and RBP-Trunc. These results exhibit a similar behavior as observed in the lysine-arginine-ornithine (LAO) binding protein (Vergara et al., 2023) but differ from those of a previous study of the type-II PBP protein HisJ (Chu et al., 2013) where the isolated lobes do not interact with each other in the presence or absence of histidine, suggesting that in HisJ only one lobe is important for ligand binding and the other is considered to play a supporting role in the dynamics of binding and in protein stability.

The differences in T_m and ΔH of the native proteins and the mixed heterodimer can be explained by the carry-over of ribose from the purification. It is notoriously hard to remove bound ligands from the expression medium when purifying solute-binding proteins that have a high affinity for their ligands (Structural Genomics Consortium et al., 2008). Due to its high stability and irreversible thermal unfolding, RBP resisted all attempts of refolding, making purification of a sample removed of all residual ribose not possible, and for this reason always some ribose was carried-over in the purified RBP, increasing the measured T_m and ΔH by a ligand stabilization mechanism. Since the individual halves of RBP do not show any binding of ribose (Figure S3), carry-over is not expected to occur during

TABLE 1 Characterization summary (oligomeric state and thermostability with/without ribose) for the constructs analyzed in this work.

Protein			Oligomeric state	T_m (°C) (protein)	T_m (°C) (protein + ribose)	Ribose binding[a]
First generation	RBP		Monomer	106.9 ± 0.4	114.0 ± 0.9	Yes
	RBP-N		Monomer	76.6 ± 0.2	76.7 ± 0.3	No
	RBP-Trunc		Monomer (90%) Homodimer (10%)	73.3 ± 0.1	73.4 ± 0.2	No
	RBP-N/RBP-Trunc mixed heterodimer		Dimer of heterodimers	99.7 ± 0.3	113.5 ± 0.4	Yes
Second generation	RBP-$N_{N\text{-}His}$		Monomer	73.2 ± 0.1	73.1 ± 0.2	No
	RBP-$TruncII_{N\text{-}Strep}$		Monomer	70.6 ± 0.2	70.8 ± 0.3	No
	RBP-$TruncII_{N\text{-}His}$		Monomer	70.4 ± 0.4	70.9 ± 0.5	No
	RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ co-expressed heterodimer		Heterodimer	104.8 ± 0.3	113.9 ± 0.4	Yes
	RBP-$N_{N\text{-}His}$/RBP-$C_{N\text{-}Strep}$ co-expressed heterodimer		n.d.	68.4 ± 0.5	83.5 ± 0.9	Yes

[a]Interaction with ribose was determined by changes in thermostability (T_m) and enthalpy (ΔH) parameters comparing DSC endotherms collected without and with 0.5 mM ribose.

purification, therefore no additional stabilizing effect of ribose binding is expected.

Next, we determined the crystal structure of the RBP-N/RBP-Trunc heterodimer (PDB ID: 7PU4) (Figure 4 and Table S4). The two halves indeed reconstitute the canonical RBP fold with high structural similarity, showing a Cα-RMSD of 0.41 Å of the heterodimer to the previously reported structure of unliganded RBP (PDB ID: 2FN9), confirming the aforementioned spectroscopic and calorimetric results. The heterodimer displays the same opening and twisting angle as the paternal protein, an important indicator of a native-like configuration of the heterodimer. The asymmetric unit of the crystal structure shows a dimer of RBP-N/RBP-Trunc heterodimers (Figure S5), which is in agreement with the oligomeric state observed in SEC-MALS experiments (Figure 2); however, further analysis is needed to determine the precise conformation of the dimer of heterodimers in solution. The observed heterodimer interface in the asymmetric unit is mostly related to the interaction of C-terminal residues of RBP-Trunc located in the hinge region and their corresponding ones in the crystallography mates, ruling out the possibility that dimerization results from the extra elements left out in RPB-Trunc. Finally, a closer look at the side-chains involved in ribose binding reveals an almost identical orientation compared

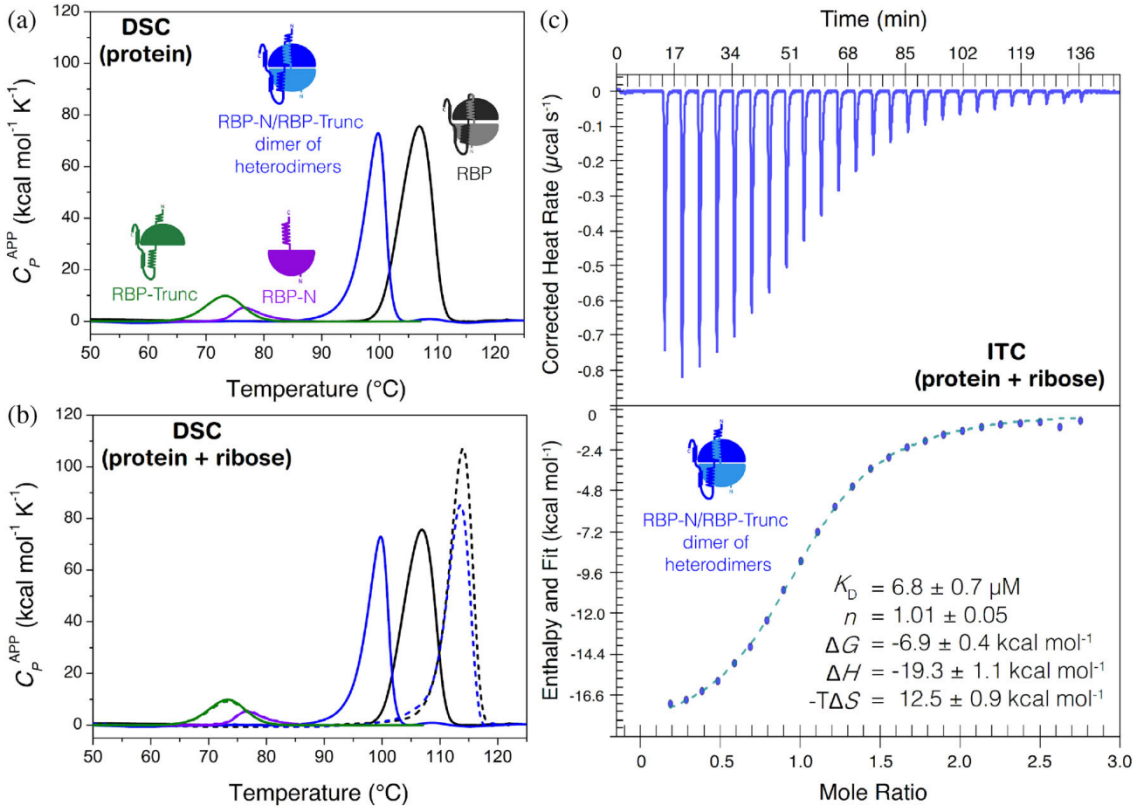


FIGURE 3 Thermodynamic characterization of the first-generation constructs and their interaction with ribose. (a) DSC endotherms at 1.5°C min^{-1} of the halves RBP-Trunc (green), RBP-N (violet), the RBP-N/RBP-Trunc dimer of heterodimers (blue), and the full-length RBP (black) without ribose and (b) with 0.5 mM ribose. Experiments were performed in 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8, and the physical and chemical baselines have been subtracted. (c) Representative ITC measurement for ribose binding of the RBP-N/RBP-Trunc heterodimer. Baseline-subtracted raw data are shown at the top while the binding isotherms (blue circles) fitted to a 1:1 model (dotted line) are presented at the bottom. ± at the reported parameters indicate the standard deviation of 3 independent experiments. Titrations were performed at 20°C in 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8.

to the unliganded state of the native RBP, suggesting the correct formation of the preformed binding site (Figure 4b). Since all these results showed that the separately purified RBP halves can reassemble the structural conformation of full-length RBP *in vitro,* we next wanted to determine whether this RBP-N/RBP-Trunc heterodimer is also a functional RBP protein.

2.4 | The reassembled heterodimer binds ribose with a comparable affinity to full-length RBP

The structural similarity of the heterodimer with the full-length RBP suggests that also the ribose binding function might be reconstituted. To investigate this, we first analyzed by DSC if ribose binding increases protein thermostability. Specific protein-ligand interaction commonly causes an increase in protein thermostability, which is due to the coupling between binding and unfolding processes under thermodynamic equilibrium (Cooper et al., 2000; Privalov, 1979).

The isolated RBP-N and RBP-Trunc do not show any sign of stabilization upon addition of ligand (Figure S3). This differs from type II PBP-like fold proteins such as LAO, ArgBP, or HisJ, in which it has been shown that albeit with lower affinity, one of the isolated lobes is able to bind its respective ligand (Chu et al., 2013; Smaldone et al., 2020; Vergara et al., 2023).

In type I PBP-like fold proteins like RBP, the binding residues are distributed almost equally between the two lobes, while in many type II PBPs almost all binding residues are present only in one lobe (mostly in the discontinuous one). In addition, changes in the hinge connections between the distinct types of PBPs also modify the binding properties and dynamics (Bermejo et al., 2009, 2010; Chu et al., 2013; Gouridis et al., 2021; Ortega et al., 2012; Pistolesi et al., 2011). These differences in the architecture of type I and II PBP-like fold proteins could explain why one of the isolated lobes from

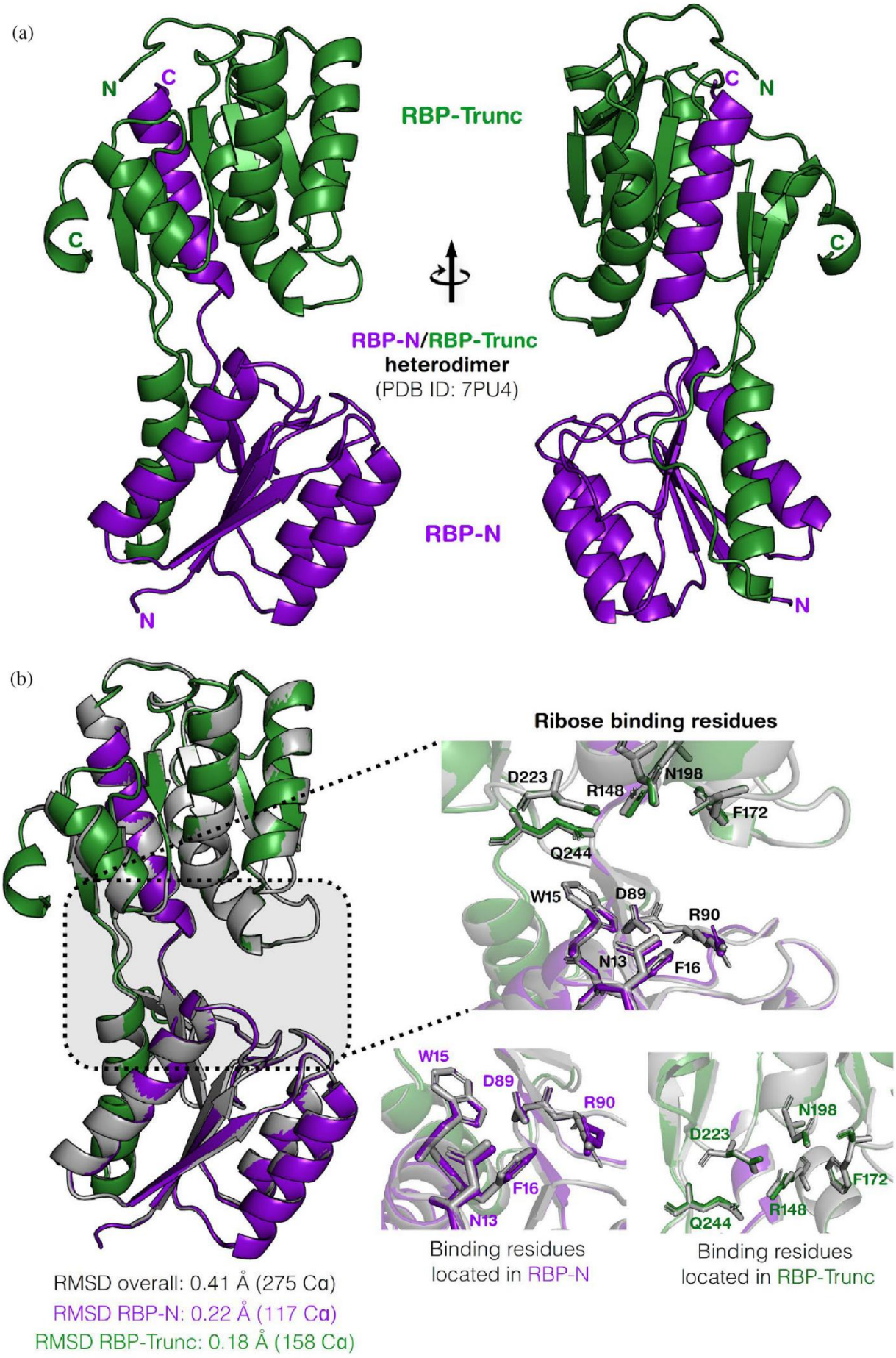


FIGURE 4 Legend on next page.

type II proteins such as LAO, HisJ, and ArgBP is able to bind their respective ligand while none of the individual lobes from type I PBPs have been shown to be competent by themselves. Variations in ligand affinity and promiscuity for some of the studied PBPs (Chu et al., 2013; Kröger, Shanmugaratnam, Ferruz, et al., 2021; Kröger, Shanmugaratnam, Scheib, et al., 2021; Vergara et al., 2020) indicate that possibly the PBP ancestor was able to bind some ligands but with considerably lower affinity, similarly to what has been reported for enzyme evolution (Copley, 2020; Khersonsky & Tawfik, 2010; Tawfik, 2020). In a plausible scenario, after duplication and fusion of the flavodoxin-like fold ancestor (Figure 1a), type I PBP-like fold proteins were able to evolve obtaining increased selectivity and affinity for specific compounds but still sharing almost equally the ligand binding residues between both domains, as has been observed for RBP.

In contrast to the isolated RBP domains, an increase in T_m can be observed upon addition of 0.5 mM ribose (Figure 3b) to the RBP-N/RBP-Trunc heterodimer, with the amplitude of the absorbed heat changes being dependent on ligand concentration. The T_m of the ligand-bound RBP-N/RBP-Trunc heterodimer increases by almost 14°C from 99.7 to 113.5°C, comparable to the stabilization of ligand-bound RBP by around 7°C to 114.0°C (Figure S3 and Table S3) and similar to the one observed in other PBPs when binding their respective high-affinity ligands (Fukada et al., 1983; Ganesh et al., 1997; Kreimer et al., 2000).

In addition, an increase of 129 kcal mol^{-1} was observed in the unfolding ΔH for the ligand-bound RBP-N/RBP-Trunc heterodimer in comparison to the unbound form, deducing that large-scale rearrangements in the solvent-exposed surface in the heterodimer accompanies ligand binding, thereby confirming a functional protein that behaves similar to full-length RBP. The greater amount of thermostabilization in the heterodimer in comparison to RBP can again be explained by residual ribose carried over in the purification of RBP already stabilizing the protein. However, at the same concentration of ribose the level of stabilization of the heterodimer is almost identical to that of RBP, with the heterodimer displaying a native-like thermostability. Interestingly, the significant increase in stability can also be observed when adding ribose to a non-native SDS-PAGE. At concentration of 1 mM ribose or higher, a dimer (and higher oligomers) can be detected, indicating that the addition of SDS and the subsequent heating to 99°C is not enough to dissociate the ribose-bound stabilized heterodimer (Figure S4).

Additionally, to DSC analysis, ribose binding of the RBP-N/RBP-Trunc dimer was determined by ITC. Ribose-binding isotherms (Figure 3c) showed a sigmoidal profile with the ribose binding constant ($K_D = 6.8 \pm 0.7$ μM) in a concentration range comparable to other previously studied solute-binding proteins (Schreier et al., 2009), implying that the binding of ribose can be regained after *in vitro* mixing the previously dissected RBP halves. In fact, ligand affinity is not significantly affected by the assembly. Now the question remained, whether this reassembled functional heterodimer can also be formed *in vivo* upon co-expression of both halves.

2.5 | RBP halves form a functional heterodimer when co-expressed in *E. coli*

To investigate whether the heterodimer of RBP-N and RBP-Trunc already forms during the expression in *E. coli*, a second generation of constructs was created (Table S1). To ensure that at least one plasmid copy of each construct stays in each cell, the coding sequences were assembled in a vector imparting resistance to either ampicillin or kanamycin, respectively. Since there was no control of expression levels and we wanted to only obtain heterodimer in the subsequent purification, we opted for adding two different affinity tags to each construct (Figure 5a). The resulting constructs are RBP-$N_{N\text{-}His}$ and RBP-Trunc$II_{N\text{-}Strep}$ (Table S1) with affinity labels located at the N-terminus. By utilizing a three-step purification approach using the different affinity tags on each protein half and a subsequent SEC step for polishing, we can assure that only already formed heterodimers are retained as confirmed by the SDS-PAGE showing a band at the corresponding sizes of both RBP-$N_{N\text{-}His}$ and RBP-Trunc$II_{N\text{-}Strep}$ and thermal resistance upon addition of ribose (Figure 5b). Similar to the behavior of the 1st generation constructs, the far-UV CD and fluorescence spectra showed a reconstitution of characteristics almost identical to the native RBP (Figure S2B and Figure S6A). The molecular weight determined by SEC-MALS also corresponds to the heterodimer (expected mass:

FIGURE 4 Crystal structure of the RBP-N/RBP-Trunc heterodimer in unliganded conformation. (a) Cartoon representation of RBP-N (violet) and RBP-Trunc (green) heterodimer (PDB ID: 7PU4) forming a native-like conformation as full-length RBP. (b) Structural comparison of RBP-N/RBP-Trunc heterodimer and RBP (PDB ID: 2FN9; gray). RMSD values are reported for the entire heterodimer, and halves RBP-N and RBP-Trunc. Inset shows the ribose binding residues in the full structure (top) and separated in each half (bottom); numbering is based on the RBP sequence.

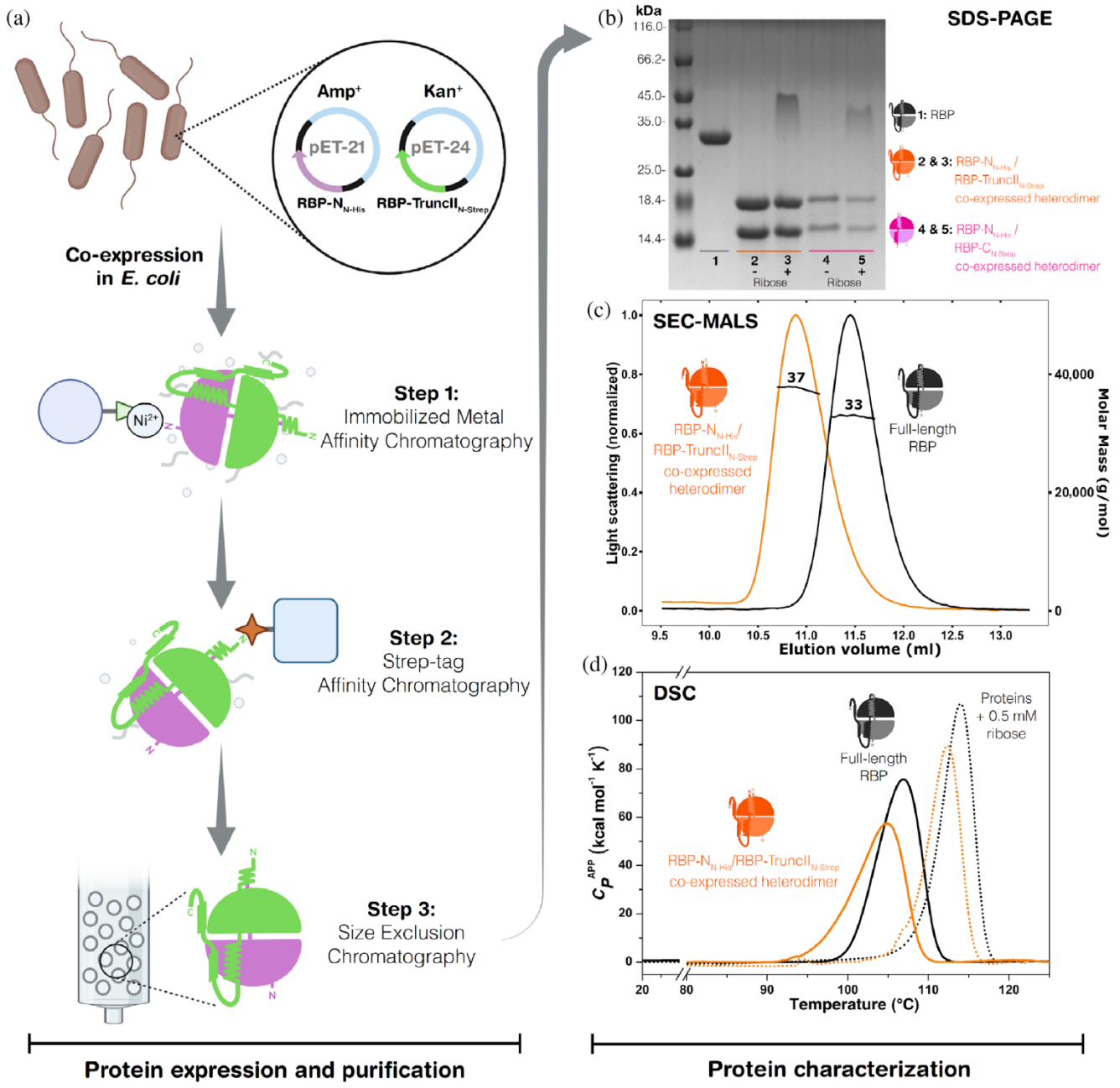


FIGURE 5 Co-expression in *Escherichia coli* and characterization of the second-generation heterodimers. (a) Schematic workflow of the co-expression beginning with the transformation of *E. coli* with the two plasmids carrying RBP-$N_{N\text{-}His}$ and RBP-$TruncII_{N\text{-}Strep}$. Subsequent alternating affinity chromatographies utilizing two different tags assure purification of only the RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ heterodimer, followed by a final size exclusion step. (b) SDS-PAGE showing the co-purified heterodimers. RBP (lane 1), co-expressed RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ heterodimer without ribose (lane 2) and with 0.5 mM ribose (lane 3), co-expressed RBP-$N_{N\text{-}His}$/RBP-$C_{N\text{-}Strep}$ heterodimer without ribose (lane 4) and with 0.5 mM ribose (lane 5). (c) SEC-MALS measurements of RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ heterodimer (orange line) in comparison with full-length RBP (black line). Numbers indicate the determined experimental molecular weight. (d) DSC endotherms of RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ heterodimer and RBP in absence (continuous lines) and presence of 0.5 mM ribose (dotted lines).

36.8 kDa/determined mass: 37.3 kDa), with no higher oligomers present (Figure 5c and Table S2).

Similar to the mixed RBP-N/RBP-Trunc heterodimer, the co-expressed and co-purified heterodimer shows an increase in thermostability in the presence of ribose (Figure 5d and Table 1), indicating a functional heterodimer. The T_m of RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ increases by 9.1°C (from 104.8 to 113.9°C) after addition of 0.5 mM

ribose, showing a similar trend of stabilization as the full-length RBP (Figure S3 and Table S3) and also to the 1st generation of halves. In view of the heterodimer being formed in the cells during co-expression, the same behavior of carrying over residual ribose from *E. coli* is expected to increase the measurable T_m of the RBP-$N_{N\text{-}His}$/RBP-Trunc$II_{N\text{-}Strep}$ heterodimer.

Since the formation of the heterodimer appears to stabilize the individual protein halves and yields properties almost identical to full-length RBP, we set out to retry the expression of the previously insolubly expressing RBP-C in the hopes that the co-expression and formation of the heterodimer *in-cell* could rescue the protein. The second generation RBP-$C_{N\text{-}Strep}$ was purified along RBP-$N_{N\text{-}His}$ analogously to the previous co-expression assay (Figure 5a). Interestingly, we were able to obtain a small amount of purified heterodimer after the affinity chromatography and subsequent SEC (Figure 5b), with a high-oligomer band still being visible in the SDS-PAGE after addition of ribose, which indicates retention of binding function. This characteristic is confirmed by DSC measurements of the heterodimer with and without ribose (Figure S7). While the overall transition is massively decreased for the unbound proteins ($T_m = 68.4°C$ for RBP-$N_{N\text{-}His}$/RBP-$C_{N\text{-}His}$ co-expressed heterodimer versus $T_m = 106.9°C$ for full-length RBP), the strong stabilization after addition of ribose is still observed (15.1°C of T_m increase to 83.5°C). The total shift is comparable with that in RBP, albeit with some fraction of the protein still appearing to be in a ligand-free state (Figure S7), indicating a possible reduction in ribose binding affinity or different populations of the purified heterodimer. The ability of RBP-$N_{N\text{-}His}$ to recover not just the soluble expression of RBP-$C_{N\text{-}Strep}$ via the formation of the heterodimer, but also the heterodimer to retain its function, showcases the inherent versatility of this fold and gives insights into its evolution.

The PBP architecture, like multidomain proteins, illustrates how the modular reuse of domains can generate more complex macromolecules, that often include the addition of extra secondary structural elements or even larger decorations towards acquiring new functions (Das et al., 2015; Ferruz et al., 2020; Gouridis et al., 2021). In a global manner, these changes have shown how domain-domain interactions, previously not present in the single independent units, are essential for the folding, stability, and function of multidomain proteins, especially for those residues located close to the interdomain interface (Vogel et al., 2004) and for modulation of binding-site solvation (Vergara et al., 2020; Vergara et al., 2023). Stabilizing interdomain interactions are useful to avoid misfolding and aggregation in multidomain proteins (Han et al., 2007) and moreover, domain-domain interactions can control the dynamics and kinetics between open and closed states, being critical factors for the transport rate of PBPs (Gouridis et al., 2015). This suggests that after the duplication and fusion of an ancestral protein that corresponded to an individual RBP lobe, the entire protein sequence now works as an integrated functional unit, where folding, stability, and binding function are interlinked. This allows the protein to evolve new properties such as gaining ligand selectivity, increasing binding affinity, and modifying the dynamics of ligand binding and transport by including open and closed states. These significant closing and twisting motions observed in PBPs (Chu et al., 2013; Gouridis et al., 2021; Kröger, Shanmugaratnam, Ferruz, et al., 2021; Vergara et al., 2020) would not be possible without the evolution of RBP as a single functional unit.

3 | CONCLUSIONS

3.1 | Implications for the evolution of the PBP fold and protein engineering approaches

The data presented here shows how a modern PBP can be disassembled into its two lobes, and how when they are combined *in vitro* or *in vivo* the formed dimer is able to perform its original function. The individual parts readily assemble to form a heterodimer, not just when mixing the individually purified lobes, but also within the cell upon co-expression. While the N- and C-terminal lobes appear to be stable and well-behaved proteins on their own, formation of the heterodimer almost completely restores the characteristics of the full-length RBP, confirming the importance of interdomain interactions on the evolution, stability and function of the PBP fold, similarly to what has been reported for other multidomain proteins (Alvarez-Carreño et al., 2022; Han et al., 2007; Vogel et al., 2004).

Analysis of the stability and binding abilities indicate native-like properties, and the crystal structure of the heterodimer being nearly identical to that reported for RBP supports this conclusion. This versatility of the PBP fold can be explained by the inherent malleability of proteins of the flavodoxin-like (and related) folds. Several structures with swapped elements have been reported for flavodoxin-like proteins (e.g., PDBs: 4Q37, 6ER7/6EXR, 3C85; Paithankar et al., 2019; Farías-Rico et al., 2014) as well as TIM-barrel proteins (PDB 6QKY; Michalska et al., 2020), which are also thought to be related to the flavodoxin-like fold (Romero-Romero et al., 2021). Further, we had previously observed swapped elements in circular-permuted constructs of RBP (PDBs: 7QSP, 7QSQ; Michel et al., 2023). This tendency of the structural archetype to enable formation of swapped

elements could have been an important characteristic promoting the emergence of the ancestral dimer thought to be the progenitor of modern PBPs. While the two halves we describe in this work are derived from an already evolved protein, they could still be seen as a vestige of this ancestral dimer. Interestingly, the crystal structure of a flavodoxin-like fold protein with an identical arrangement of secondary structure elements has been described already, albeit it is unclear whether the observed structure is an artifact of the non-physiological crystallographic conditions (Lewis et al., 2000).

Since the heterodimer corresponds to the proposed ancestral dimer in the evolutionary trajectory (Figure 1a) while still retaining function with native-like properties, this presents new insight into the mechanisms behind such a duplication event. Not only does the orientation of the two lobes create the binding cleft characteristic for PBP-like proteins, but also the general restraints on the movement of the lobes lower the entropic cost of ligand binding. Our findings showcase the feasibility of a functional heterodimer similar to the proposed ancestral one to also assemble within cells, giving way to the argument that the duplication and fusion of the progenitor flavodoxin-like protein might have happened independent of the gain of function, indicating no evolutive pressure on single domains but on the full-length RBP.

Adopting this approach and expanding it to incorporate a diverse set of functions could also be used for protein engineering purposes. This is traditionally done by inserting a domain for readout into the sequence of an existing PBP, with the optimal placement of the insertion sites being one of the major challenges (Ribeiro et al., 2019; Tullman et al., 2016). Further studies will have to show that the retracing of the duplication is applicable for other PBPs as well, but one could imagine its usage in creating modular switch systems not just *in vitro*, but also *in vivo*.

4 | MATERIALS AND METHODS

4.1 | Reagents and solutions

Analytical grade chemicals were used for all the experiments. Water was distilled and deionized.

4.2 | Identification of the protein halves and sequence analysis

The bioinformatic analysis to trace the sequence similarities between the RBP and flavodoxin-like proteins was done using the HHpred server which is part of the HHsuite (Gabler et al., 2020) (Figure S1). The sequence of full-length RBP (UniProt-ID: Q9X053) excluding the extracellular transport signal was run with standard parameters, but disabling secondary structure scoring and increasing the number of maximal hits to 10,000 to also obtain sequences with lower probability scores. Based on the alignment of both the other PBP lobes and the hits with the flavodoxin-like proteins, the cutting points were determined at position 30–155 for RBP-N, 142–310 for RBP-Trunc, 156–310 for RBP-TruncII, and 157–291 for RBP-C (Table S1).

4.3 | Cloning and generation of RBP-constructs

The gene fragment for wild-type RBP lacking the periplasmic signal sequence as well as the primers used for assembly were provided by Eurofins Genomics. To generate the gene fragments for RBP-N and RBP-Trunc, a polymerase chain-reaction with the corresponding primer was conducted with the full sequence as template. Additionally, a QuikChange® site-directed mutagenesis was performed to obtain the M142A mutation of the full-length RBP to prevent the translation of the truncated protein (henceforth called RBP). The fragment of full length RBP was cloned into empty pET-21 using the *Nde*I/*Xho*I restriction sites. Analogously generated fragments for RBP, RBP-N, and RBP-Trunc were all subsequently cloned using T5 exonuclease-dependent assembly (Xia et al., 2019). All constructs were verified by sequencing.

Gene synthesis and cloning for the co-expression assay were provided by Biocat. The differently tagged constructs of RBP-TruncII and RBP-$N_{N\text{-}His}$ were cloned into pET24- and pET21-vectors, respectively. Individual clones were obtained by transforming *E. coli* BL21 (DE3) cells by adding 50 ng of purified plasmid, heat shock and subsequent plating on agar-plates supplemented with the corresponding antibiotic. To obtain cells carrying the two different plasmids needed for the co-expression assay, 50 ng of each plasmid were added to the *E. coli* BL21 (DE3) cells, heat shocked and then grown on plates containing the two selecting antibiotics.

4.4 | Expression and purification of RBP-constructs

The transformant *E. coli* BL21(DE3) were grown in *Terrific broth* media (TB) at 37°C to an OD_{600} of 1.2 in the presence of the corresponding antibiotics (ampicillin 100 μg mL^{-1}; kanamycin 50 μg mL^{-1}). Protein expression was induced by the addition of Isopropyl-β-thiogalactopyranoside to a concentration of 1 mM and

a total time of 18 h at 20°C. Cells were harvested via centrifugation (5000 × G, 15 min), resuspended in the corresponding binding buffer (20 mL g^{-1} wet weight), lysed by sonication and subsequently centrifuged to remove remaining cell debris (40,000 × G, 1 h). The cleared lysate was filtered through a 0.22 μm filter previous to the affinity column step.

For the constructs carrying a hexahistidine affinity tag, Immobilized Metal Ion Chromatography (IMAC) was performed on a Cytiva HisTrap 5 mL column equilibrated with buffer (20 mM MOPS, 500 mM sodium chloride, 10 mM imidazole, pH 7.8). Elution was performed with a step of IMAC-Elution-Buffer (20 mM MOPS, 500 mM sodium chloride, 600 mM imidazole, pH 7.8) at 40%, and fractions corresponding to the eluted protein pooled and concentrated to a volume suitable for the size exclusion chromatography (SEC) step.

Strep-Tactin affinity chromatography was used for constructs with a StrepII-Tag, which were loaded onto a Cytiva StrepTrap HP 5 mL column equilibrated with Strep-Trap binding Buffer (100 mM Tris–HCl, 150 mM sodium chloride, 1 mM EDTA, pH 7.8) and eluted with Strep-Trap elution Buffer (100 mM Tris–HCl, 150 mM sodium chloride, 1 mM EDTA, 2.5 mM Desthiobiotin, pH 7.8), pooled and concentrated analogous to the IMAC purification. To facilitate purification of the individual constructs, the Strep-Tag of RBP-$Trunc_{N\text{-}Strep}$ was switched to a His_6-Tag, creating RBP-$Trunc_{N\text{-}His}$.

For the purification of the co-expressed constructs, to assure survival of cells carrying only the two plasmids, the LB medium used for the production was supplemented with both Ampicillin and Kanamycin (100 and 50 μg mL^{-1}, respectively). Cell lysis was performed as with the individual constructs, and the lysate first loaded on the HisTrap column. The eluted fractions corresponding to the tagged protein were pooled and applied onto a StrepTrap column. Similarly, eluted fractions were pooled and concentrated to a volume suitable for application onto the Superdex column.

SEC was performed as final purification step for all constructs on a Cytiva Superdex 26/600 75 pg with an isocratic elution using buffer 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8. Fractions consistent with the proteins of interest were analyzed by SDS-PAGE, pooled, flash frozen in liquid nitrogen, and stored at −20°C until further analysis.

4.5 | Far-UV circular dichroism

Far-UV circular dichroism (CD) measurements were performed at 20°C in buffer 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8 in a Jasco J-710 spectropolarimeter equipped with a Peltier device to control temperature (PTC-348 WI). Spectra were collected using 5 μM protein concentration for RBP and the heterodimers, and 10 μM for the other constructs in a 2 mm cuvette, 195–260 nm wavelength range, and 1 nm bandwidth. After buffer subtraction, raw data were converted to mean residue molar ellipticity ([Θ]) with $[\Theta] = \Theta/l\ C\ N_r$, where Θ is the ellipticity signal in millidegrees, l is the cell path in mm, C is the molar protein concentration, and N_r is the number of amino acids per protein (Greenfield, 2006).

4.6 | Intrinsic fluorescence

Intrinsic fluorescence (IF) spectra were collected on a Jasco FP-6500 spectrofluorometer coupled with a water bath (Julabo MB) to control the temperature. Experiments were performed at 20°C in buffer 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8, and 5 μM protein concentration for RBP and heterodimers, and 5 μM for the other proteins, with 280 nm as excitation wavelength, 300–500 nm as emission wavelength, and 1 nm bandwidth. Raw signal was normalized for protein concentration.

4.7 | Analytical size exclusion chromatography coupled to multi angle light scattering (SEC-MALS)

Analytical SEC measurements were performed coupled to a miniDAWN Multi Angle Light Scattering (MALS) detector and an Optilab refractometer (Wyatt Technology). Samples previously centrifuged and filtered were run in a Superdex 75 Increase 10/300 GL column connected to an Äkta Pure System (GE Healthcare Life Sciences) equilibrated with buffer 10 mM sodium phosphate, 50 mM sodium chloride, 0.02% sodium azide, pH 7.8. Experiments were conducted at room temperature with a protein concentration of 1 and 0.8 mL min^{-1} flow rate. For the samples containing ribose, 0.5 mM of ribose was premixed with protein at 1 mg mL^{-1}. Reproducibility during all SEC-MALS measurements was tested by running a BSA standard at 2 mg mL^{-1} at the beginning and end of all experiments, which resulted in identical data. Determination of weight averaged molar mass was performed by using the Zimm-Equation with the differential refractive index signal as source for the concentration calculations (refractive index increment dn/dc set to 0.185). Data collection and analysis were done using the ASTRA v.7.3.2 software (Wyatt Technology).

4.8 | Crystallization and three-dimensional structure determination

For setting up crystallization assays, protein at 0.5 mM concentration was dialyzed against 20 mM Tris–HCl, 300 mM sodium chloride, pH 7.8. For RBP-N/RBP-Trunc heterodimer, 0.5 mM equimolar ratio of each protein was used as initial concentration. Screening plates were set up by a sitting-drop vapor diffusion method using JCSG Core I-IV (Qiagen), PEG Suite I-II (Qiagen), and Additive Screen kits (Hampton Research) in 96 well Intelli plates (Art Robbins Instruments). Plates with 0.8 μL drops in a 1:1, 1:2, and 2:1 protein: mother liquor drop ratio were set up with a nano dispensing crystallization Phoenix robot (Art Robbins Instruments) and stored at 20°C in a hotel-based crystal imaging system RockImager RI 1000 (Formulatrix). RBP-N/RBP-Trunc heterodimer crystals with successful diffraction data were found in 100 mM HEPES pH 7.5, 15% (w/v) PEG 20000 and a drop ratio 1:1. Data were collected at Berlin Electron Storage Ring Society for Synchrotron Radiation beamline 14.2 (BESSY 14.2) operated by the Helmholtz-Zentrum Berlin using the mxCuBE beamline-control software (Gabadinho et al., 2010). Measurements at 100 K were performed in a single-wavelength mode at 0.9184 Å with a PILATUS3S 2M detector (HZB, 2016) in fine-slicing mode (0.1° wedges). Diffraction images were processed with x-ray detector software (XDS) and XDSAPP v3.0 (Kabsch, 2010; Sparta et al., 2016). Phasing was performed by molecular replacement with PHASER in the PHENIX software suite v.1.19.2 (Liebschner et al., 2019) using the edited pdb file corresponding to the RBP-N and RBP-Trunc halves from *T. maritima* RBP (PDB 2FN9). Data refinement was carried out with phenix.refine (Adams et al., 2010) and iterative manual model building/improvement in COOT v.0.9 (Emsley et al., 2010). Coordinates and structure factors were validated and deposited in the PDB database https://www.rcsb.org/ (Berman et al., 2002) with the accession code: 7PU4. Figures were created with PyMOL Molecular Graphics System v.2.3.0 (Schrodinger, LLC).

4.9 | Differential scanning calorimetry

Differential scanning calorimetry (DSC) endotherms were collected using a VP-Capillary DSC instrument (Malvern Panalytical) with a temperature range of 10–130°C and 1.5°C min^{-1} scan rate. Protein samples were prepared at 50 μM after exhaustive dialysis in buffer 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8, and proper degassing. Instrument equilibration was performed by collecting at least two buffer–buffer scans before each protein-buffer experiment. Calorimetric reversibility was tested by collecting two consecutive endotherms and calculating the recovery area percentage from the second and first scan, resulting in irreversible thermal-unfolding transitions for all the constructs reported in the present study. Thermodynamic parameters (T_m and ΔH) were calculated after subtracting physical (buffer–buffer scan) and chemical baselines (heat capacity effects) from each protein-buffer scan. Thermostabilization by protein–protein interaction (dimer formation) was determined by changes in T_m and ΔH when two different proteins were combined in equimolar concentration. DSC experiments in presence of ribose were performed at 50 μM protein concentration and 0.5 mM ribose premixed in the same working buffer before the heating cycles. Buffer–buffer scans were collected containing the same amount of ribose as protein/ribose-buffer experiments and subtracted as indicated. Ribose stability at high temperatures was tested and no endotherm distortions were observed in the concentration and temperature ranges assayed. Origin v.7.0 (OriginLab Corporation) with MicroCal software was used for data analysis.

4.10 | Isothermal titration calorimetry

Binding assays followed by isothermal titration calorimetry (ITC) were performed using a TA Nano ITC low volume device (TA Instruments). Titrations were obtained at 20°C in buffer 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8, and 100 μM of protein concentration, which was exhaustively dialyzed against the working buffer. Ribose solution was prepared in the same working buffer to minimize dilution heats and was loaded in the syringe at 0.8 mM concentration. Protein and ligand solutions were degassed with a vacuum pump for 90 min before carrying out the experiments, and concentrations were optimized in order to reach c values higher than 10. Independent triplicates of ITC experiments were performed with 25 injections of 2 μL volume, spacing of 350 s between injections, and stirring at 300 rpm. Dilution heats were subtracted from the heat associated with each injection to get accurate parameters. Baseline and integration intervals were carefully checked to avoid experiment distortions. Binding constant (K_D), enthalpy change (ΔH), and binding stoichiometry (n) were determined by nonlinear fitting of normalized data assuming a 1:1 binding model and using TA ITC software. All titration replicates fulfilled the characteristics for an accurate parameter determination that have been analyzed by experimental and simulation data (Turnbull & Daranas, 2003).

AUTHOR CONTRIBUTIONS

Florian Michel, Sergio Romero-Romero, and Birte Höcker designed the research, Florian Michel, Sergio

Romero-Romero purified the different constructs, Florian Michel collected CD, IF, and SEC-MALS data, Sergio Romero-Romero performed DSC and ITC experiments, Florian Michel, Sergio Romero-Romero crystallized and solved three-dimensional structure, Florian Michel, Sergio Romero-Romero, Birte Höcker analyzed the date and wrote the manuscript.

ACKNOWLEDGMENTS

We acknowledge allocation of synchrotron beamtime and financial support by HZB and thank the beamline staff at BESSY for support. We thank Saacnicteh Toledo-Patiño for scientific discussions and input in the early stages of the project, Sabrina Wischt and Sooruban Shanmugaratnam for their competent technical support, and all the members of the Höcker Lab for their constructive suggestions to improve the research. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

This work was supported by the European Research Council (ERC Consolidator Grant 647548 "Protein Lego" to Birte Höcker), the VolkswagenStiftung (grant 94747 to Birte Höcker), and by a fellowship from the Alexander von Humboldt and Bayer Science & Education Foundation (Humboldt-Bayer Research Fellowship for Postdoctoral Researchers to Sergio Romero-Romero).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest with the contents of this article.

DATA AVAILABILITY STATEMENT

All data to support the conclusions of this manuscript are included in the main text and Supporting Information. Coordinates and structure factors have been deposited to the Protein Data Bank (PDB) with accession code: 7PU4 (RBP-N/RBP-Trunc heterodimer).

ORCID

Florian Michel https://orcid.org/0000-0002-5111-8290
Sergio Romero-Romero https://orcid.org/0000-0003-2144-7912
Birte Höcker https://orcid.org/0000-0002-8250-9462

REFERENCES

Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: A comprehensive python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66(2):213–21. https://doi.org/10.1107/S0907444909052925

Aggarwal V, Kulothungan SR, Balamurali MM, Saranya SR, Varadarajan R, Ainavarapu SR. Ligand-modulated parallel mechanical unfolding pathways of maltose-binding proteins. J Biol Chem. 2011;286(32):28056–65. https://doi.org/10.1074/jbc.M111.249045

Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. elife. 2015;4:e09410. https://doi.org/10.7554/eLife.09410

Alvarez-Carreño C, Gupta RJ, Petrov AS, Williams LD. Creative destruction: new protein folds from old. Proc Natl Acad Sci USA. 2022;119(52):e2207897119. https://doi.org/10.1073/pnas.2207897119

Bae JE, Kim IJ, Kim KJ, Nam KH. Crystal structure of a substrate-binding protein from Rhodothermus marinus reveals a single α/β-domain. Biochem Biophys Res Commun. 2018;497(1):368–73. https://doi.org/10.1016/j.bbrc.2018.02.086

Banda-Vázquez J, Shanmugaratnam S, Rodríguez-Sotres R, Torres-Larios A, Höcker B, Sosa-Peinado A. Redesign of LAOBP to bind novel l-amino acid ligands. Protein Sci. 2018;27(5):957–68. https://doi.org/10.1002/pro.3403

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 2002;58(6 I):899–907. https://doi.org/10.1107/S0907444902003451

Bermejo GA, Strub MP, Ho C, Tjandra N. Determination of the solution-bound conformation of an amino acid binding protein by NMR paramagnetic relaxation enhancement: use of a single flexible paramagnetic probe with improved estimation of its sampling space. J Am Chem Soc. 2009;131(27):9532–7. https://doi.org/10.1021/ja902436g

Bermejo GA, Strub MP, Ho C, Tjandra N. Ligand-free open-closed transitions of periplasmic binding proteins: the case of glutamine-binding protein. Biochemistry. 2010;49(9):1893–902. https://doi.org/10.1021/bi902045p

Berntsson RP, Smits SH, Schmitt L, Slotboom DJ, Poolman B. A structural classification of substrate-binding proteins. FEBS Lett. 2010;584(12):2606–17. https://doi.org/10.1016/j.febslet.2010.04.043

Brandts JF, Hu CQ, Lin LN, Mos MT. A simple model for proteins with interacting domains. Applications to scanning calorimetry data. Biochemistry. 1989;28(21):8588–96. https://doi.org/10.1021/bi00447a048

Careaga CL, Sutherland J, Sabeti J, Falke JJ. Large amplitude twisting motions of an interdomain hinge: a disulfide trapping study of the galactose-glucose binding protein. Biochemistry. 1995;34(9):3048–55. https://doi.org/10.1021/bi00009a036

Chandonia JM, Fox NK, Brenner SE. SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. Nucleic Acids Res. 2019;47(D1):D475–81. https://doi.org/10.1093/nar/gky1134

Chandravanshi M, Tripathi SK, Kanaujia SP. An updated classification and mechanistic insights into ligand binding of the substrate-binding proteins. FEBS Lett. 2021;595(18):2395–409. https://doi.org/10.1002/1873-3468.14174

Cheng H, Liao Y, Schaeffer RD, Grishin NV. Manual classification strategies in the ECOD database. Proteins. 2015;83(7):1238–51. https://doi.org/10.1002/prot.24818

Chu BC, DeWolf T, Vogel HJ. Role of the two structural domains from the periplasmic *Escherichia coli* histidine-binding protein HisJ. J Biol Chem. 2013;288(44):31409–22. https://doi.org/10.1074/jbc.M113.490441

Clifton BE, Jackson CJ. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. Cell Chem Biol. 2016;23(2):236–45. https://doi.org/10.1016/j.chembiol.2015.12.010

Cooper A, Nutley MA, Walood A. In: Harding SE, Chowdhry BZ, editors. Differential scanning microcalorimetry. *Oxford, New York: Oxford University Press*; 2000. p. 287–318.

Copley SD. Evolution of new enzymes by gene duplication and divergence. FEBS J. 2020;287(7):1262–83. https://doi.org/10.1111/febs.15299

Cuneo MJ, Beese LS, Hellinga HW. Ligand-induced conformational changes in a thermophilic ribose-binding protein. BMC Struct Biol. 2008;8:50. https://doi.org/10.1186/1472-6807-8-50

Das S, Dawson NL, Orengo CA. Diversity in protein domain superfamilies. Curr Opin Genet Dev. 2015;35:40–9. https://doi.org/10.1016/j.gde.2015.09.005

Dwyer MA, Hellinga HW. Periplasmic binding proteins: a versatile superfamily for protein engineering. Curr Opin Struct Biol. 2004;14(4):495–504. https://doi.org/10.1016/j.sbi.2004.07.004

Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 4):486–501. https://doi.org/10.1107/S0907444910007493

Farías-Rico JA, Schmidt S, Höcker B. Evolutionary relationship of two ancient protein superfolds. Nat Chem Biol. 2014;10(9):710–5. https://doi.org/10.1038/nchembio.1579

Felder CB, Graul RC, Lee AY, Merkle HP, Sadee W. The Venus flytrap of periplasmic binding proteins: an ancient protein module present in multiple drug receptors. AAPS PharmSci. 1999;1(2):E2–E26. https://doi.org/10.1208/ps010202

Ferruz N, Lobos F, Lemm D, Toledo-Patino S, Farías-Rico JA, Schmidt S, et al. Identification and analysis of natural building blocks for evolution-guided fragment-based protein design. J Mol Biol. 2020;432(13):3898–914. https://doi.org/10.1016/j.jmb.2020.04.013

Ferruz N, Michel F, Lobos F, Schmidt S, Höcker B. Fuzzle 2.0: Ligand binding in natural protein building blocks. Front Mol Biosci. 2021;8:715–972. https://doi.org/10.3389/fmolb.2021.715972

Fukada H, Sturtevant JM, Quiocho FA. Thermodynamics of the binding of L-arabinose and of D-galactose to the L-arabinose-binding protein of *Escherichia coli*. J Biol Chem. 1983;258(21):13193–8. https://doi.org/10.1016/S0021-9258(17)44100-7

Fukami-Kobayashi K, Tateno Y, Nishikawa K. Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history. J Mol Biol. 1999;286(1):279–90. https://doi.org/10.1006/jmbi.1998.2454

Gabadinho J, Beteva A, Guijarro M, Rey-Bakaikoa V, Spruce D, Bowler MW, et al. MxCuBE: a synchrotron beamline control environment customized for macromolecular crystallography experiments. J Synchrotron Radiat. 2010;17(5):700–7. https://doi.org/10.1107/S0909049510020005

Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein sequence analysis using the MPI bioinformatics toolkit. Curr Protoc Bioinformatics. 2020;72(1):e108. https://doi.org/10.1002/cpbi.108

Ganesh C, Shah AN, Swaminathan CP, Surolia A, Varadarajan R. Thermodynamic characterization of the reversible, two-state unfolding of maltose binding protein, a large two-domain protein. Biochemistry. 1997;36(16):5020–8. https://doi.org/10.1021/bi961967b

Gouridis G, Muthahari YA, de Boer M, Griffith DA, Tsirigotaki A, Tassis K, et al. Structural dynamics in the evolution of a bilobed protein scaffold. Proc Natl Acad Sci U S A. 2021;118(49):e2026165118. https://doi.org/10.1073/pnas.2026165118

Gouridis G, Schuurman-Wolters GK, Ploetz E, Husada F, Vietrov R, de Boer M, et al. Conformational dynamics in substrate-binding domains influences transport in the ABC importer GlnPQ. Nat Struct Mol Biol. 2015;22(1):57–64. https://doi.org/10.1038/nsmb.2929

Greenfield NJ. Using circular dichroism spectra to estimate protein secondary structure. Nat Protoc. 2006;1(6):2876–90. https://doi.org/10.1038/nprot.2006.202

Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J. The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol. 2007;8(4):319–30. https://doi.org/10.1038/nrm2144

Helmholtz-Zentrum Berlin für Materialien und Energie. The MX beamlines BL14.1–3 at BESSY II. JLSRF. 2016;2:A47. https://doi.org/10.17815/jlsrf-2-64

Jeffery CJ. Engineering periplasmic ligand binding proteins as glucose nanosensors. Nano Rev. 2011;2:5743. https://doi.org/10.3402/nano.v2i0.5743

Kabsch W. XDS. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):125–32. https://doi.org/10.1107/S0907444909047337

Kantaev R, Riven I, Goldenzweig A, Barak Y, Dym O, Peleg Y, et al. Manipulating the folding landscape of a multidomain protein. J Phys Chem B. 2018;122(49):11030–8. https://doi.org/10.1021/acs.jpcb.8b04834

Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem. 2010;79:471–505. https://doi.org/10.1146/annurev-biochem-030409-143718

Kreimer DI, Malak H, Lakowicz JR, Trakhanov S, Villar E, Shnyrov VL. Thermodynamics and dynamics of histidine-binding protein, the water-soluble receptor of histidine permease. Implications for the transport of high and low affinity ligands. Eur J Biochem. 2000;267(13):4242–52. https://doi.org/10.1046/j.1432-1033.2000.01470.x

Kröger P, Shanmugaratnam S, Ferruz N, Schweimer K, Höcker B. A comprehensive binding study illustrates ligand recognition in the periplasmic binding protein PotF. Structure. 2021;29(5):433–43.e4. https://doi.org/10.1016/j.str.2020.12.005

Kröger P, Shanmugaratnam S, Scheib U, Höcker B. Fine-tuning spermidine binding modes in the putrescine binding protein PotF. J Biol Chem. 2021;297(6):101419. https://doi.org/10.1016/j.jbc.2021.101419

Lewis RJ, Muchová K, Brannigan JA, Barák I, Leonard G, Wilkinson AJ. Domain swapping in the sporulation response regulator SpoOA. J Mol Biol. 2000;297(3):757–70. https://doi.org/10.1006/jmbi.2000.3598

Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, et al. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Biol Crystallogr. 2019;75(Pt 10):861–77. https://doi.org/10.1107/S2059798319011471

Liu K, Chen, X, & Kaiser, CM (2019). Energetic dependencies dictate folding mechanism in a complex protein. Proceedings of

the National Academy of Sciences of the United States of America, 116(51), 25641–8. https://doi.org/10.1073/pnas.1914366116

Louie GV. Porphobilinogen deaminase and its structural similarity to the bidomain binding proteins. Curr Opin Struct Biol. 1993; 3:401–8. https://doi.org/10.1016/S0959-440X(05)80113-7

Matilla MA, Ortega Á, Krell T. The role of solute binding proteins in signal transduction. Comput Struct Biotechnol J. 2021;19: 1786–805. https://doi.org/10.1016/j.csbj.2021.03.029

Medintz IL, Deschamps JR. Maltose-binding protein: a versatile platform for prototyping biosensing. Curr Opin Biotechnol. 2006;17(1):17–27. https://doi.org/10.1016/j.copbio.2006.01.002

Michalska K, Kowiel M, Bigelow L, Endres M, Gilski M, Jaskolski M, et al. 3D domain swapping in the TIM barrel of the α subunit of Streptococcus pneumoniae tryptophan synthase. Acta Crystallogr D Biol Crystallogr. 2020;76(Pt 2): 166–75. https://doi.org/10.1107/S2059798320000212

Michel F, Shanmugaratnam S, Romero-Romero S, Höcker B. Structures of permuted halves of a modern ribose-binding protein. Acta Crystallogr D Biol Crystallogr. 2023;79(Pt 1):40–9. https://doi.org/10.1107/S205979832201186X

Nepomnyachiy S, Ben-Tal N, Kolodny R. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. Proc Natl Acad Sci U S A. 2017;114(44):11703–8. https://doi.org/10.1073/pnas.1707642114

Ortega G, Castaño D, Diercks T, Millet O. Carbohydrate affinity for the glucose-galactose binding protein is regulated by allosteric domain motions. J Am Chem Soc. 2012;134(48):19869–76. https://doi.org/10.1021/ja3092938

Paithankar KS, Enderle M, Wirthensohn DC, Miller A, Schlesner M, Pfeiffer F, et al. Structure of the archaeal chemotaxis protein CheY in a domain-swapped dimeric conformation. Acta Crystallogr F: Struct Biol Commun. 2019;75, (Pt 9):576–85. https://doi.org/10.1107/S2053230X19010896

Pistolesi S, Tjandra N, Bermejo GA. Solution NMR studies of periplasmic binding proteins and their interaction partners. Biomol Concepts. 2011;2(1–2):53–64. https://doi.org/10.1515/bmc.2011.005

Prajapati RS, Indu S, Varadarajan R. Identification and thermodynamic characterization of molten globule states of periplasmic binding proteins. Biochemistry. 2007;46(36):10339–52. https://doi.org/10.1021/bi700577m

Privalov PL. Stability of proteins: small globular proteins. Adv Protein Chem. 1979;33:167–241. https://doi.org/10.1016/s0065-3233(08)60460-x

Ribeiro LF, Amarelle V, Ribeiro LFC, Guazzaroni ME. Converting a periplasmic binding protein into a synthetic biosensing switch through domain insertion. BioMed Res Int. 2019;2019:4798793 10.1155/2019/4798793.

Romero-Romero S, Kordes S, Michel F, Höcker B. Evolution, folding, and design of TIM barrels and related proteins. Curr Opin Struct Biol. 2021;68:94–104. https://doi.org/10.1016/j.sbi.2020.12.007

Scheepers GH, Nijeholt JALA, Poolman B. An updated structural classification of substrate-binding proteins. FEBS Lett. 2016;590(23):4393–401. https://doi.org/10.1002/1873-3468.12445

Scheib U, Shanmugaratnam S, Farías-Rico JA, Höcker B. Change in protein-ligand specificity through binding pocket grafting. J Struct Biol. 2014;185(2):186–92. https://doi.org/10.1016/j.jsb.2013.06.002

Schreier B, Stumpp C, Wiesner S, Höcker B. Computational design of ligand binding is not a solved problem. Proc Natl Acad Sci U S A. 2009;106(44):18491–6. https://doi.org/10.1073/pnas.0907950106

Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res. 2021;49(D1):D266–73. https://doi.org/10.1093/nar/gkaa1079

Smaldone G, Ruggiero A, Balasco N, Vitagliano L. Development of a protein scaffold for arginine sensing generated through the dissection of the arginine-binding protein from *Thermotoga maritima*. Int J Mol Sci. 2020;21(20):7503. https://doi.org/10.3390/ijms21207503

Sparta KM, Krug M, Heinemann U, Mueller U, Weiss MS. XDSAPP2.0. J Appl Crystallogr. 2016;49:1085–92. https://doi.org/10.1107/S1600576716004416

Steffen V, Otten J, Engelmann S, Radek A, Limberg M, Koenig BW, et al. A toolbox of genetically encoded FRET-based biosensors for rapid l-lysine analysis. Sensors. 2016;16(10):1604. https://doi.org/10.3390/s16101604

Structural Genomics Consortium, China Structural Genomics Consortium, Northeast Structural Genomics Consortium, Gräslund S, Nordlund P, Weigelt J, et al. Protein production and purification. Nat Methods. 2008;5(2):135–46. https://doi.org/10.1038/nmeth.f.202

Tawfik DS. Enzyme promiscuity and evolution in light of cellular metabolism. FEBS J. 2020;287(7):1260–1. https://doi.org/10.1111/febs.15296

Toledo-Patiño S, Chaubey M, Coles M, Höcker B. Reconstructing the remote origins of a fold singleton from a Flavodoxin-like ancestor. Biochemistry. 2019;58(48):4790–3. https://doi.org/10.1021/acs.biochem.9b00900

Tullman J, Nicholes N, Dumont MR, Ribeiro LF, Ostermeier M. Enzymatic protein switches built from paralogous input domains. Biotechnol Bioeng. 2016;113(4):852–8. https://doi.org/10.1002/bit.25852

Turnbull WB, Daranas AH. On the value of c: can low affinity systems be studied by isothermal titration calorimetry? J Am Chem Soc. 2003;125(48):14859–66. https://doi.org/10.1021/ja036166s

Vergara R, Berrocal T, Juárez Mejía EI, Romero-Romero S, Velázquez-López I, Pulido NO, et al. Thermodynamic and kinetic analysis of the LAO binding protein and its isolated domains reveal non-additivity in stability, folding and function. FEBS J. 2023;290:4496–512. https://doi.org/10.1111/febs.16819

Vergara R, Romero-Romero S, Velázquez-López I, Espinoza-Pérez G, Rodríguez-Hernández A, Pulido NO, et al. The interplay of protein-ligand and water-mediated interactions shape affinity and selectivity in the LAO binding protein. FEBS J. 2020;287(4):763–82. https://doi.org/10.1111/febs.15019

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol. 2004;14(2):208–16. https://doi.org/10.1016/j.sbi.2004.03.011

Wenk M, Jaenicke R, Mayr EM. Kinetic stabilisation of a modular protein by domain interactions. FEBS Lett. 1998;438(1–2):127–30. https://doi.org/10.1016/s0014-5793(98)01287-3

Xia Y, Li K, Li J, Wang T, Gu L, Xun L. T5 exonuclease-dependent assembly offers a low-cost method for efficient cloning and site-directed mutagenesis. Nucleic Acids Res. 2019;47(3):e15. https://doi.org/10.1093/nar/gky1169

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Michel F, Romero-Romero S, Höcker B. Retracing the evolution of a modern periplasmic binding protein. Protein Science. 2023;32(11):e4793. https://doi.org/10.1002/pro.4793

Supplementary Information for:

Retracing the evolution of a modern periplasmic binding protein

Florian Michel,[1†] Sergio Romero-Romero,[1†] Birte Höcker[1*]

[1] Department of Biochemistry, University of Bayreuth, Bayreuth 95447, Germany.

[†] These authors contributed equally to the work.

Correspondence

* Corresponding author. Birte Höcker. Department of Biochemistry, University of Bayreuth, Bayreuth 95447, Germany. Phone: +490921557845. **E-mail:** birte.hoecker@uni-bayreuth.de

This file includes:

- Supplementary figures 1-7.
- Supplementary tables 1-4.

Supplementary figures and tables

List of Supplementary Figures:

- **Figure S1.** Representative HHpred results for the RBP sequence.
- **Figure S2**. Intrinsic fluorescence measurements of the first- and second-generation constructs.
- **Figure S3**. DSC experiments for the first- and second-generation constructs.
- **Figure S4**. SDS-PAGE of RBP, the individual first-generation halves, and the mixed heterodimer.
- **Figure S5**. Crystallographic dimer formed by the asymmetric-unit mate of RBP-N/RBP-Trunc heterodimer crystal structure.
- **Figure S6**. Biophysical characterization of the second-generation constructs.
- **Figure S7**. DSC endotherms for the co-expressed RBP-$N_{N\text{-}His}$/RBP-$C_{N\text{-}Strep}$ heterodimer.

List of Supplementary Tables:

- **Table S1.** Amino acid sequences of the proteins analyzed in this work.
- **Table S2.** Obtained values from the SEC-MALS measurements for the different RBP constructs.
- **Table S3.** DSC thermodynamic parameters (T_m and ΔH) for the different RBP constructs in absence and presence of ribose.
- **Table S4.** Data collection and refinement statistics for crystal structures.

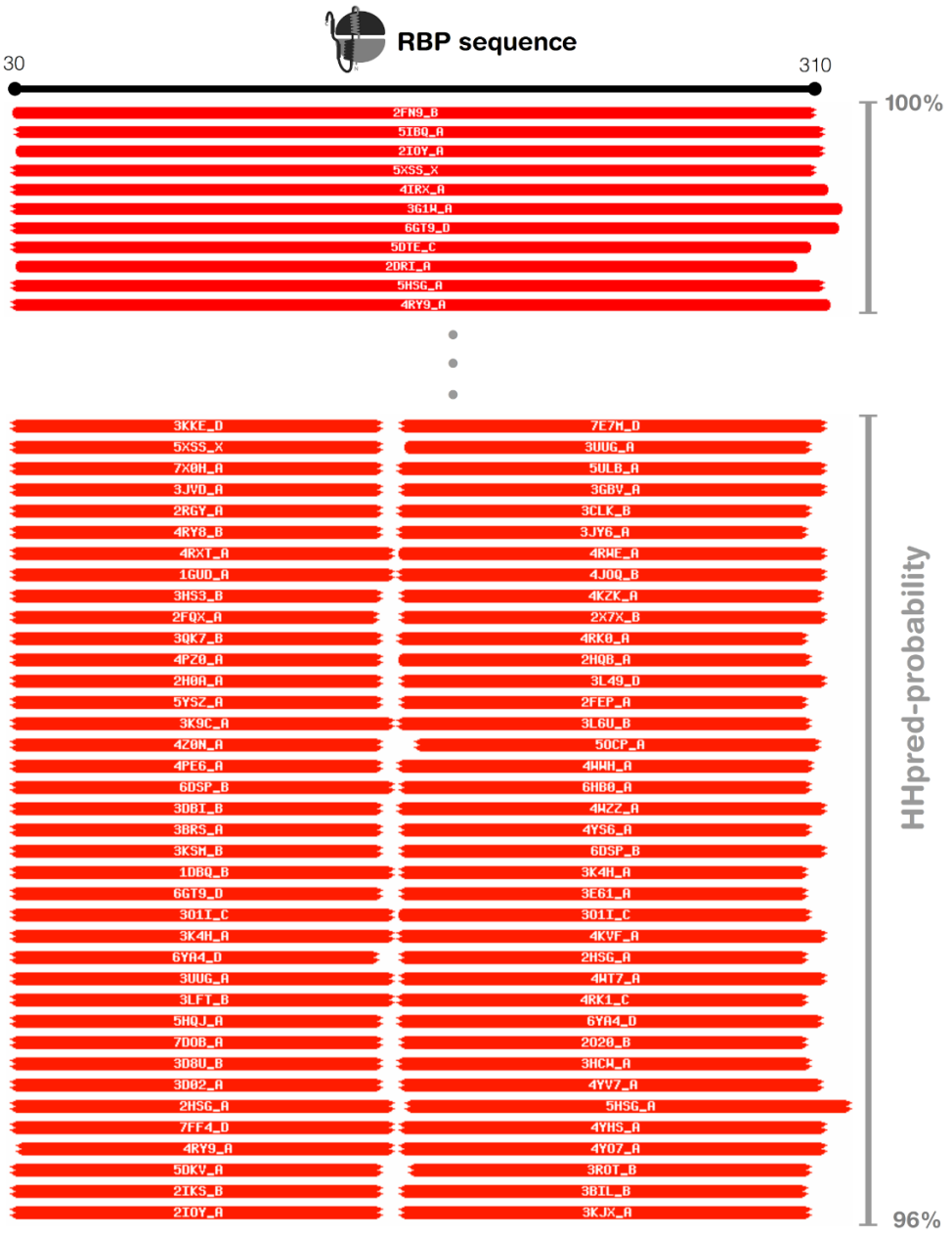


Figure S1. Representative HHpred results for the RBP sequence. Visualization of the HHpred output showing the query sequence as a black bar. The database matches are shown as red horizontal bars underneath with their respective identifiers. Bar length is indicating its coverage with respect to the query and is colored according to its significance (red as very significant to orange, yellow, green and cyan as less significant). Top and longer bars show the alignment of other full-length PBPs on the query sequence while bottom and shorter bars indicate the alignment of the individual lobes. On the right is the HHpred probability shown for the presented sequence range. Numbering has been adapted to be consistent with uniprot entry Q9X053.

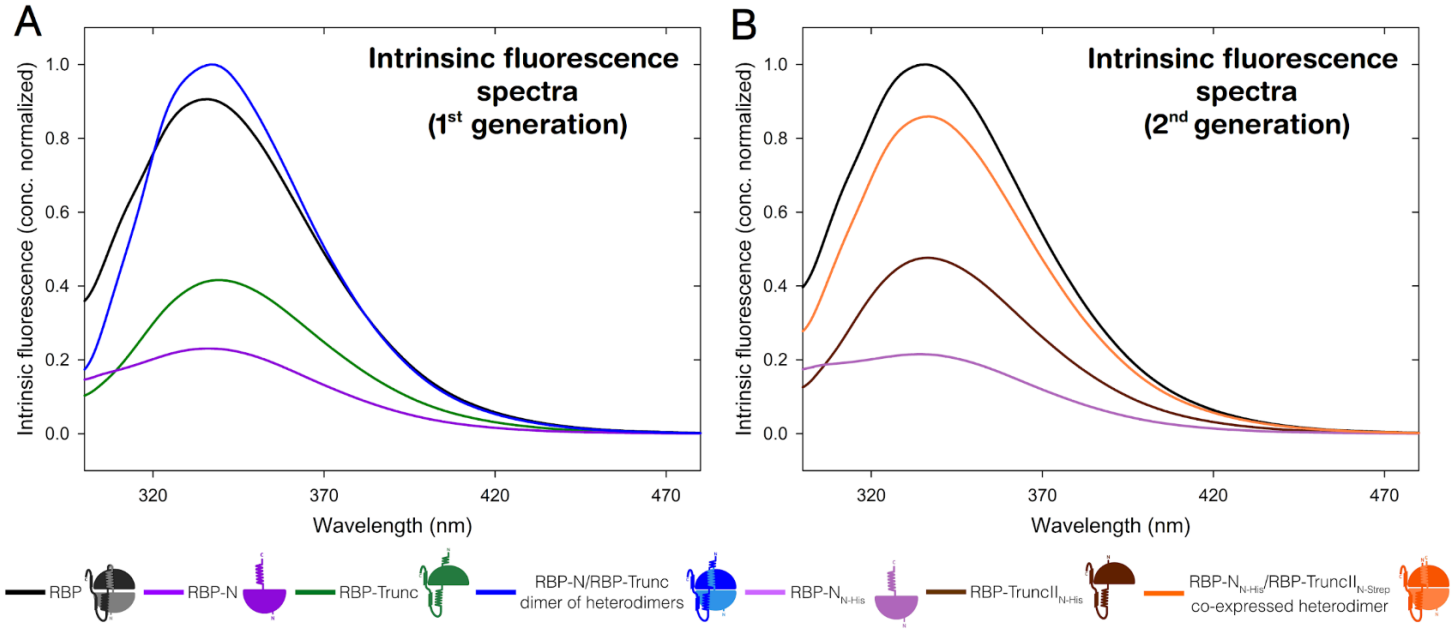


Figure S2. Intrinsic fluorescence measurements of the first- and second-generation constructs. Fluorescence spectra measured from 300-480 nm at an excitation wavelength of 280 nm of RBP (black), RBP-N (violet), RBP-Trunc (green) and the mixed RBP-N/RBP-Trunc heterodimer (**A**) and RBP-$N_{N\text{-}His}$ (red), RBP-$TruncII_{N\text{-}His}$ (brown) and the co-expressed RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ heterodimer (**B**) in 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8. Signal was normalized by protein concentration.

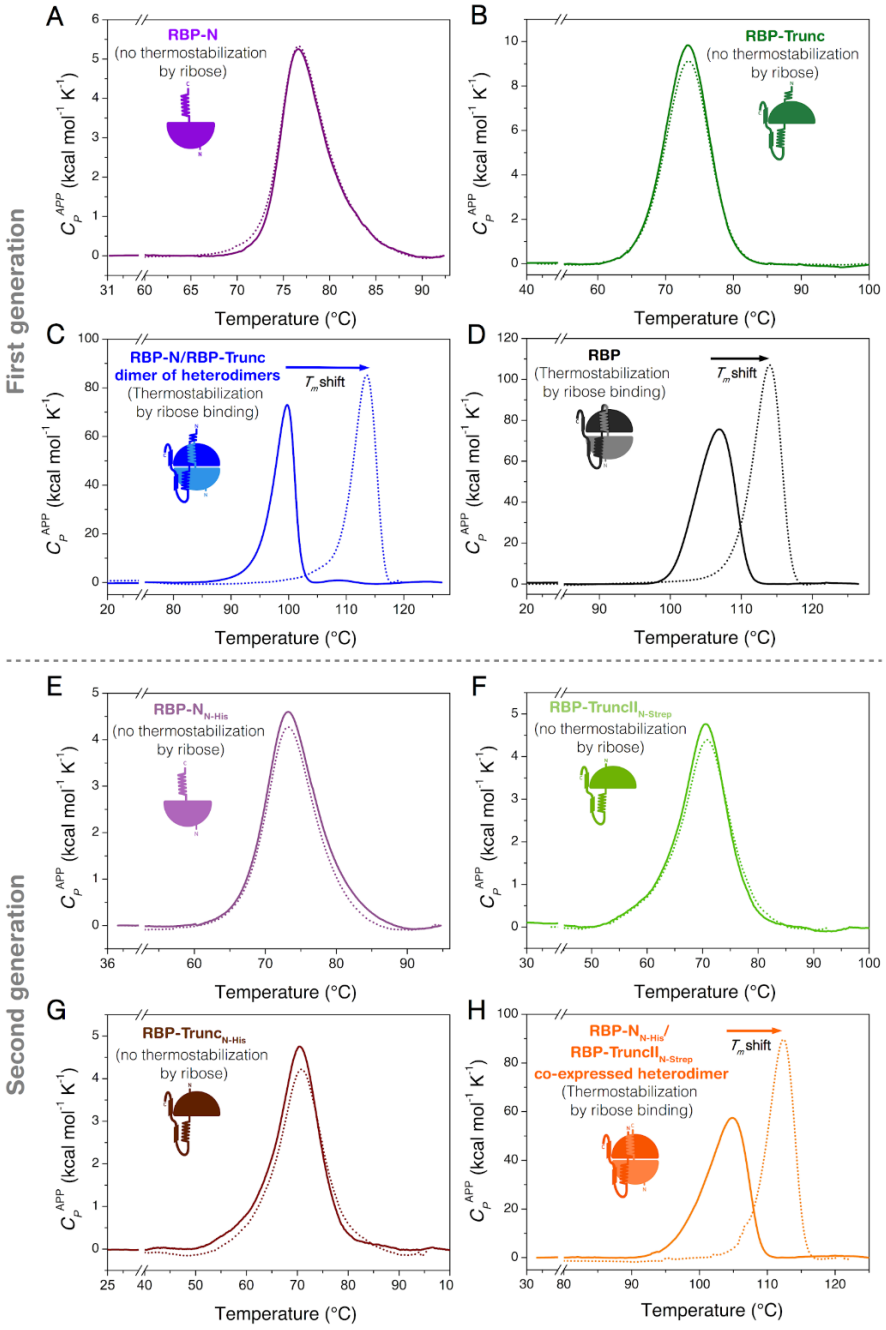


Figure S3. DSC experiments for the first- and second-generation constructs. DSC endotherms at 1.5 °C min^{-1} without ribose (solid lines) and with 0.5 mM ribose (dotted lines) of **(A)** RBP-N (violet), **(B)** RBP-Trunc (green), **(C)** RBP-N/RBP-Trunc heterodimer, **(D)** full-length RBP (black), **(E)** RBP-N$_{N\text{-}His}$ (light purple), **(F)** RBP-TruncII$_{N\text{-}Strep}$ (light green), **(G)** RBP-TruncII$_{N\text{-}His}$ (brown), and **(H)** co-expressed RBP-N$_{N\text{-}His}$/RBP-TruncII$_{N\text{-}Strep}$ heterodimer (orange). Experiments were performed in 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8 and the physical and chemical baselines have been subtracted.

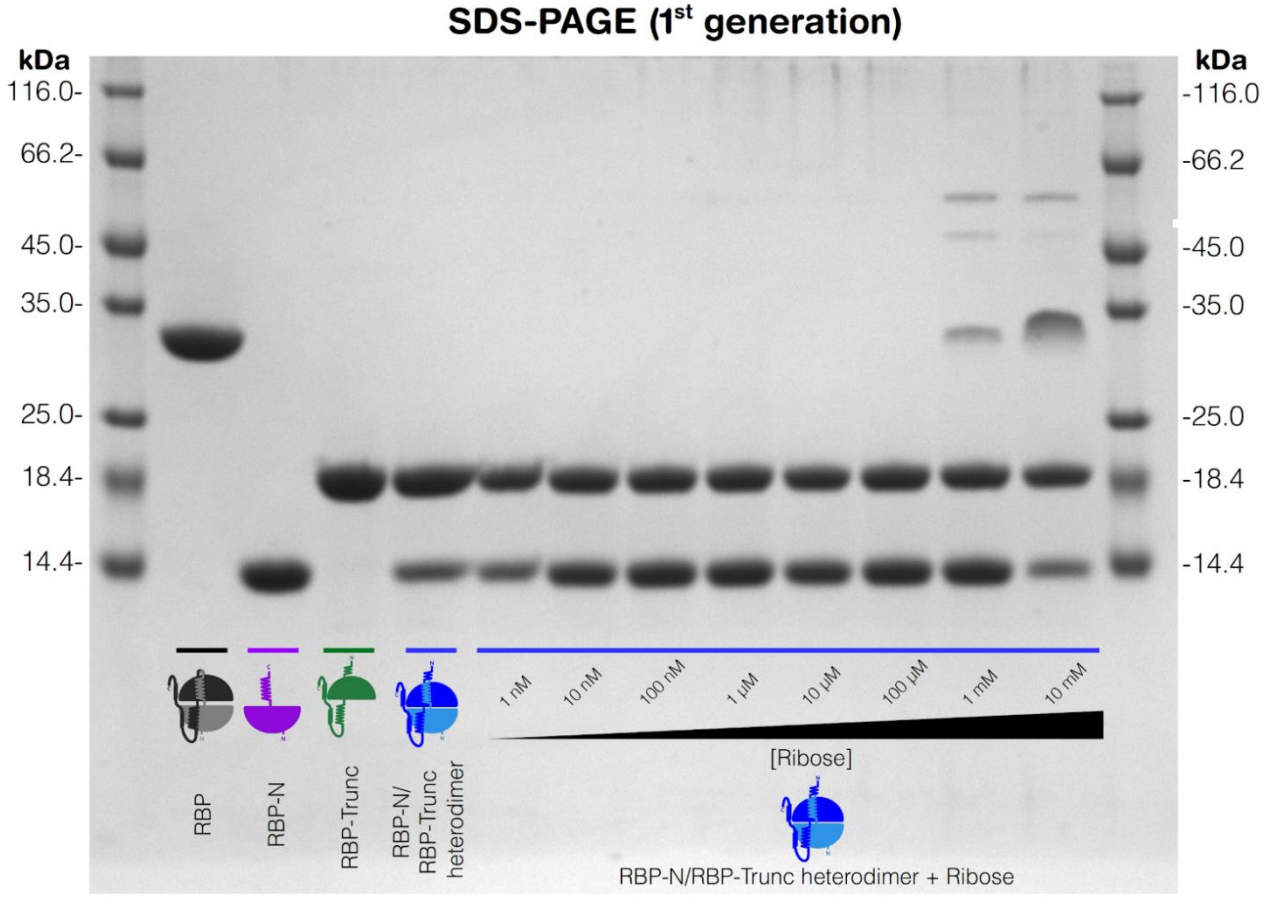


Figure S4. SDS-PAGE of RBP, the individual first-generation halves, and the mixed heterodimer. Purified RBP, RBP-N, RBP-Trunc and RBP-N/RBP-Trunc heterodimer (lane 2-5 respectively) show single proteins at the expected molecular weight without major contaminants. Addition of ribose to the heterodimer appears to stabilise the complex to a degree where it becomes resistant to dissociation in the SDS loading buffer and subsequent heating as indicated by the presence of higher oligomer bands in the presence of ≥1 mM [ribose] (lanes 6-13). Molecular weight has been estimated as indicated by the addition of the molecular weight standard (lane 1 and 14, weights annotated).

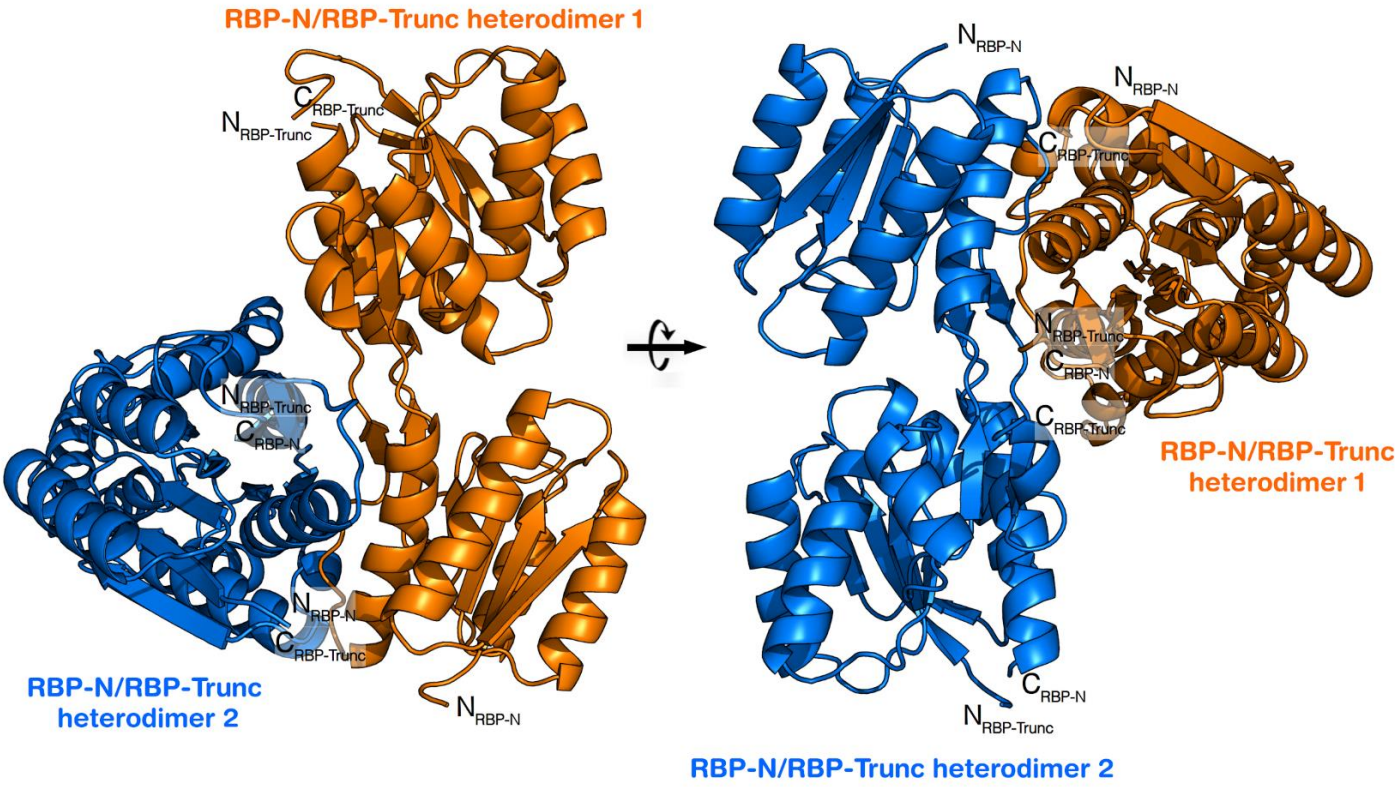


Figure S5. Crystallographic dimer formed by the asymmetric-unit mate of RBP-N/RBP-Trunc heterodimer crystal structure. Each heterodimer is indicated in orange and blue.

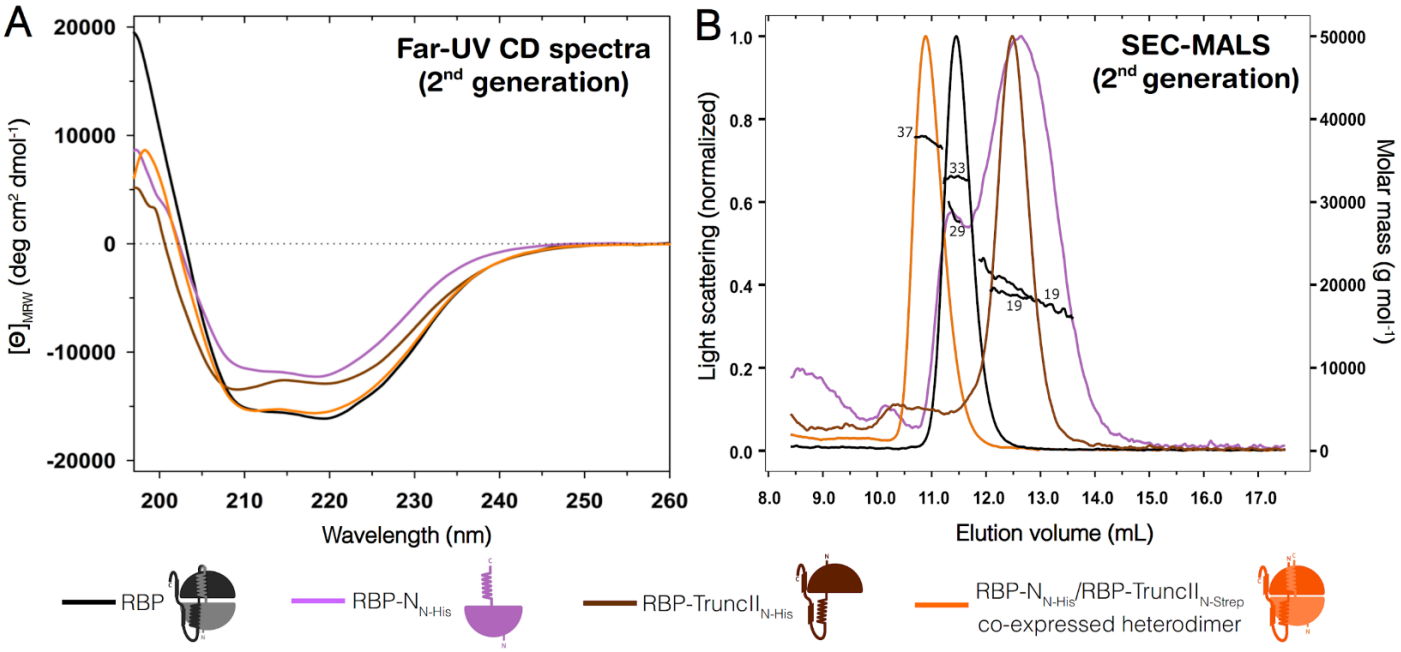


Figure S6. Biophysical characterization of the second-generation constructs. (A) Far-UV CD spectra of RBP (black), RBP-$N_{N\text{-}His}$ (light purple), RBP-$TruncII_{N\text{-}His}$ (brown) and the co-expressed RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ heterodimer (orange) in 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8. **(B)** SEC-MALS measurements in 10 mM sodium phosphate, 50 mM sodium chloride, 0.02% sodium azide, pH 7.8. Numbers indicate the determined molecular weight after data analysis. Values derived from the experiments are reported in Supplementary Table S2.

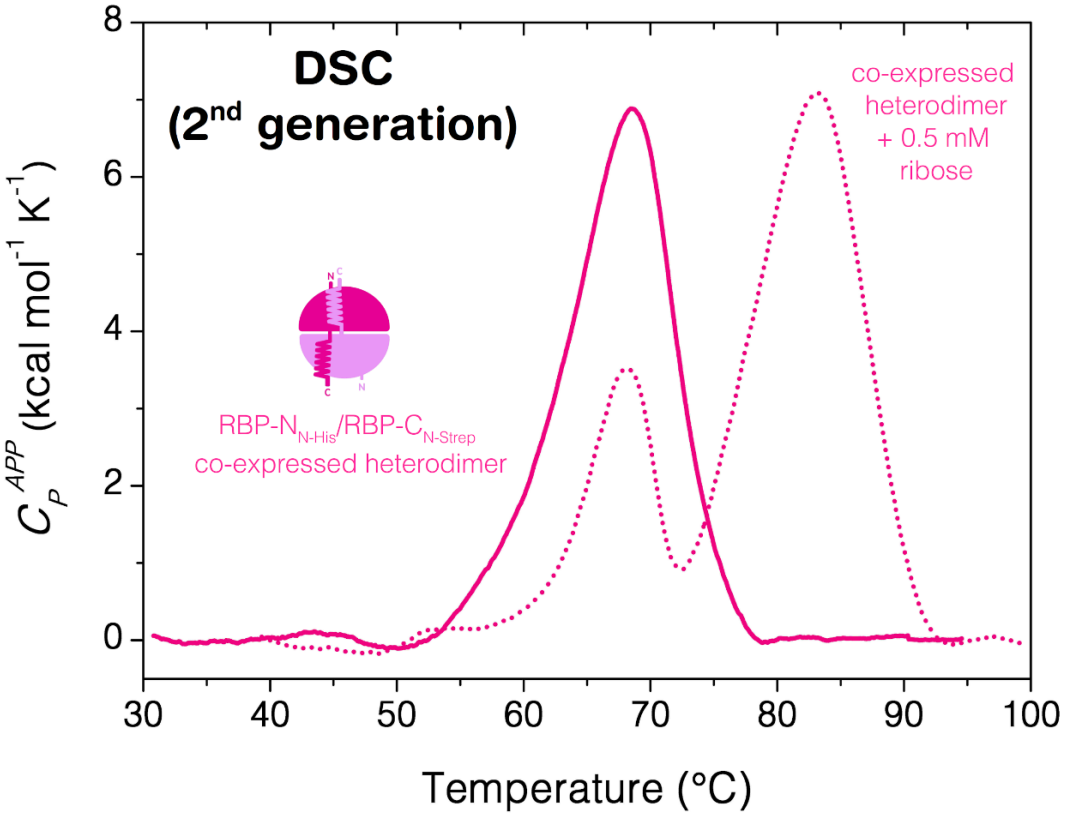


Figure S7. DSC endotherms for the co-expressed RBP-$N_{N\text{-}His}$/RBP-$C_{N\text{-}Strep}$ heterodimer. DSC experiments were collected at 1.5 °C min^{-1} without ribose (solid lines) and with 0.5 mM ribose (dotted lines) in 10 mM sodium phosphate, 50 mM sodium chloride, pH 7.8. Physical and chemical baselines were subtracted.

Table S1. Amino acid sequences of the proteins analyzed in this work. Tags used for expression/purification are highlighted in red. Differences in constructs (numbering consistent with uniprot entry Q9X053) as indicated below. RBP-N & RBP-$N_{N\text{-}His}$: correspond to the N-terminal lobe (30-153); RBP-C: corresponding to the flavodoxin-like architecture derived from the RBP C-terminal half, vestigial helix on N- and additional elements on C-terminus removed (157-291); RBP-Trunc: derived from the alternate initiation of translation at M142 (142-310); RBP-$TruncII_{N\text{-}His/N\text{-}Strep}$: corresponds to truncated construct, with the vestigial helix at the new N-terminus removed (156-310); RBP-$C_{N\text{-}Strep}$: corresponds to the C-terminal half, additional residues added of C-terminus (156-294).

	Protein	Representation	Expression tag	Sequence
First generation	RBP		His-tag (C-terminal)	MKGKMAIVISTLNNPWFVVLAETAKQRAEQLGYEATIFDSQNDTAKESAHFDAIIAAG YDAIIFNPTDADGSIANVKRAKEAGIPVFCVDRGINARGLAVAQIYSDNYYGGVLAGE YFVKFLKEKYPDAKEIPYAELLGILSAQPTWDRSNGFHSVVDQYPEFKMVAQQSAEFD RDTAYKVTEQILQAHPEIKAIWCGNDAMALGAMKACEAAGRTDIYIFGFDGAEDVINA IKEGKQIVATIMQFPKLMARLAVEWADQYLRGERSFPEIVPVTVELVTRENIDKYTAY GRK**LEHHHHHH**
	RBP-N		His-tag (C-terminal)	MKGKMAIVISTLNNPWFVVLAETAKQRAEQLGYEATIFDSQNDTAKESAHFDAIIAAG YDAIIFNPTDADGSIANVKRAKEAGIPVFCVDRGINARGLAVAQIYSDNYYGGVLMGE YFVKFLKEK**LEHHHHHH**
	RBP-C		His-tag (C-terminal)	MKEIPYAELLGILSAQPTWDRSNGFHSVVDQYPEFKMVAQQSAEFDRDTAYKVTEQIL QAHPEIKAIWCGNDAMALGAMKACEAAGRTDIYIFGFDGAEDVINAIKEGKQIVATIM QFPKLMARLAVEWADQYLR**LEHHHHHH**
	RBP-Trunc		His-tag (C-terminal)	MGEYFVKFLKEKYPDAKEIPYAELLGILSAQPTWDRSNGFHSVVDQYPEFKMVAQQSA EFDRDTAYKVTEQILQAHPEIKAIWCGNDAMALGAMKACEAAGRTDIYIFGFDGAEDV INAIKEGKQIVATIMQFPKLMARLAVEWADQYLRGERSFPEIVPVTVELVTRENIDKY TAYGRK**LEHHHHHH**
Second generation	RBP-$N_{N\text{-}His}$		His-tag (N-terminal + TEV site)	**MHHHHHHGLENLYFQGLE**KGKMAIVISTLNNPWFVVLAETAKQRAEQLGYEATIFDSQ NDTAKESAHFDAIIAAGYDAIIFNPTDADGSIANVKRAKEAGIPVFCVDRGINARGLA VAQIYSDNYYGGVLMGEYFVKFLKEK
	RBP-$C_{N\text{-}Strep}$		Strep-tag (N-terminal + TEV site)	**MWSHPQFEKGLENLYFQGLE**DAKEIPYAELLGILSAQPTWDRSNGFHSVVDQYPEFKM VAQQSAEFDRDTAYKVTEQILQAHPEIKAIWCGNDAMALGAMKACEAAGRTDIYIFGF DGAEDVINAIKEGKQIVATIMQFPKLMARLAVEWADQYLRGER
	RBP-$TruncII_{N\text{-}Strep}$		Strep-tag (N-terminal + TEV site)	**MWSHPQFEKGLENLYFQGLE**DAKEIPYAELLGILSAQPTWDRSNGFHSVVDQYPEFKM VAQQSAEFDRDTAYKVTEQILQAHPEIKAIWCGNDAMALGAMKACEAAGRTDIYIFGF DGAEDVINAIKEGKQIVATIMQFPKLMARLAVEWADQYLRGERSFPEIVPVTVELVTR ENIDKYTAYGRK
	RBP-$TruncII_{N\text{-}His}$		His-tag (N-terminal + TEV site)	**MHHHHHHGLENLYFQGLE**DAKEIPYAELLGILSAQPTWDRSNGFHSVVDQYPEFKMVA QQSAEFDRDTAYKVTEQILQAHPEIKAIWCGNDAMALGAMKACEAAGRTDIYIFGFDG AEDVINAIKEGKQIVATIMQFPKLMARLAVEWADQYLRGERSFPEIVPVTVELVTREN IDKYTAYGRK

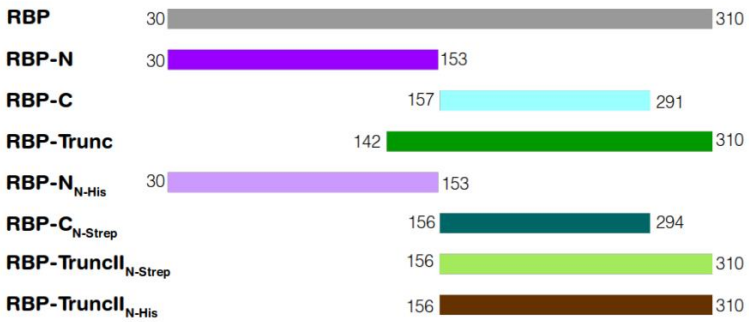


Table S2. Obtained values from the SEC-MALS measurements for the different RBP constructs.

Protein			Expected Mw (kDa)	Experimental Mw (kDa)	Polydispersity (Mw/Mn)‡	Mass Fraction (%)	Oligomeric state
Individual constructs							
RBP			33.6	32.9 ± 0.1	1.000 ± 0.005	100	Monomer
RBP + 0.5 mM ribose				32.7 ± 0.2	1.000 ± 0.008	100	Monomer
RBP-N			14.7	15.3 ± 0.2	1.001 ± 0.022	100	Monomer
RBP-Trunc	Peak 1		21.5	21.0 ± 0.2	1.000 ± 0.011	90.1	Monomer
	Peak 2			43.1 ± 0.7	1.000 ± 0.031	9.9	Homodimer
$RBP\text{-}N_{N\text{-}His}$	Peak 1		15.7	19.1 ± 0.2	1.003 ± 0.016	86.5	Monomer
	Peak 2			29.1 ± 0.5	1.002 ± 0.025	13.5	Homodimer
$RBP\text{-}TruncII_{N\text{-}His}$			20.9	18.8 ± 0.2	1.000 ± 0.015	100	Monomer
Heterodimers							
RBP-N/RBP-Trunc mixed heterodimer			36.2	69.8 ± 0.2	1.001 ± 0.005	100	Dimer of heterodimers
RBP-N/RBP-Trunc mixed heterodimer + 0.5 mM ribose	Peak 1		36.2	43.1 ± 0.2	1.001 ± 0.006	87.3	Heterodimer
	Peak 2			73.1 ± 0.7	1.000 ± 0.012	12.7	Dimer of heterodimers
$RBP\text{-}N_{N\text{-}His}$/$RBP\text{-}TruncII_{N\text{-}Strep}$ co-expressed heterodimer			36.8	37.3 ± 0.2	1.000 ± 0.007	100	Heterodimer

± indicates the standard deviation of 3 separate runs.

‡ Polydispersity was calculated by M_w/M_n; M_w - weight-average molar mass moment measured by light scattering; M_n - number-average molar mass moment. A ratio $M_w/M_n=1$ indicates a homogeneous (i.e., monodisperse) sample, because the average mass is independent of the averaging method.

Table S3. DSC thermodynamic parameters (T_m and ΔH) for the different RBP constructs in absence and presence of ribose.

Protein		T_m (°C)	ΔH (kcal mol^{-1})	Interaction with ribose‡
Individual constructs				
RBP		106.9 ± 0.4	531.4 ± 1.2	Yes
RBP + 0.5 mM ribose		114.0 ± 0.9	553.9 ± 2.0	
RBP-N		76.6 ± 0.2	35.2 ± 0.3	No
RBP-N + 0.5 mM ribose		76.7 ± 0.3	35.9 ± 0.6	
RBP-Trunc		73.3 ± 0.1	79.9 ± 0.4	No
RBP-Trunc + 0.5 mM ribose		73.4 ± 0.2	75.0 ± 1.3	
RBP-$N_{N\text{-}His}$		73.2 ± 0.1	42.9 ± 0.4	No
RBP-$N_{N\text{-}His}$ + 0.5 mM ribose		73.1 ± 0.2	38.7 ± 1.1	
RBP-$TruncII_{N\text{-}Strep}$		70.6 ± 0.2	54.4 ± 0.9	No
RBP-$TruncII_{N\text{-}Strep}$ + 0.5 mM ribose		70.8 ± 0.3	49.9 ± 1.4	
RBP-$TruncII_{N\text{-}His}$		70.4 ± 0.4	50.3 ± 0.5	No
RBP-$TruncII_{N\text{-}His}$ + 0.5 mM ribose		70.9 ± 0.5	44.9 ± 0.8	
Heterodimers				
RBP-N/RBP-Trunc mixed heterodimer		99.7 ± 0.3	355.7 ± 0.9	Yes
RBP-N/RBP-Trunc mixed heterodimer + 0.5 mM ribose		113.5 ± 0.4	484.8 ± 1.6	
RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ co-expressed heterodimer		104.8 ± 0.3	462.7 ± 2.3	Yes
RBP-$N_{N\text{-}His}$/RBP-$TruncII_{N\text{-}Strep}$ co-expressed heterodimer + 0.5 mM ribose		113.9 ± 0.4	496.1 ± 1.8	
RBP-$N_{N\text{-}His}$/RBP-$C_{N\text{-}Strep}$ co-expressed heterodimer		68.4 ± 0.5	73.9 ± 0.6	Yes
RBP-$N_{N\text{-}His}$/RBP-$C_{N\text{-}Strep}$ co-expressed heterodimer + 0.5 mM ribose		1st peak: 68.2 ± 0.7 2nd peak: 83.5 ± 0.9	1st peak: 19.2 ± 1.3 2nd peak: 83.4 ± 0.8	

‡ Interaction with ribose was determined by changes in thermostability (T_m) and enthalpy (ΔH) parameters comparing DSC endotherms collected with and without 0.5 mM ribose.

Table S4. Data collection and refinement statistics for crystal structures. Statistics for the highest resolution shell are shown in brackets.

Protein	RBP-N/RBP-Trunc heterodimer
PDB ID	7PU4
Wavelength (Å)	0.9184
Resolution range	39.01 – 1.69 (1.75 – 1.69)
Space group	P 21 21 21
Unit cell [a, b, c (Å) / α, β, γ (°)]	65.2 84.2 103.8 / 90 90 90
Total reflections	859968 (81810)
Unique reflections	64629 (6298)
Multiplicity	13.3 (12.8)
Completeness (%)	99.76 (98.39)
Mean I/sigma(I)	14.25 (0.96)
Wilson B-factor	30.6
R-merge	0.129 (1.837)
R-meas	0.127 (1.196)
R-pim	0.035 (1.104)
CC1/2	0.999 (0.318)
CC*	1.000 (0.695)
Matthews coefficient V_m (Å^3 Da^{-1})	1.97
Solvent content (%)	37.7
Protein molecules per asymmetric unit	4 halves
Reflections used in refinement	64515 (6294)
Reflections used for R-free	2095 (205)
R-work	0.206 (0.460)
R-free	0.240 (0.516)
CC(work)	0.964 (0.612)
CC(free)	0.956 (0.581)
Number of non-hydrogen atoms	4757
macromolecules	4373
ligands	0
solvent	384
Protein residues	574
RMS(bonds)	0.008
RMS(angles)	0.900
Ramachandran favored (%)	97.88
Ramachandran allowed (%)	1.77
Ramachandran outliers (%)	0.35
Rotamer outliers (%)	0.00
Clashscore	3.75
Average B-factor	43.1
macromolecules	43.0
solvent	44.6
Number of TLS groups	4

5. Paper II

Michel, F., Shanmugaratnam, S., Romero-Romero, S., Höcker, B.
Structures of permuted halves of a modern ribose-binding protein.
Acta Crystallographica Section D, 2023, D79, 40-49

research papers

STRUCTURAL BIOLOGY

ISSN 2059-7983

Structures of permuted halves of a modern ribose-binding protein

Florian Michel, Sooruban Shanmugaratnam, Sergio Romero-Romero and Birte Höcker*

Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany. *Correspondence e-mail: birte.hoecker@uni-bayreuth.de

Received 2 August 2022
Accepted 13 December 2022

Edited by P. Langan, Institut Laue-Langevin, Grenoble, France

Keywords: periplasmic binding proteins; ribose binding protein; *Thermotoga maritima*; flavodoxin-like fold; circular permutation; domain swapping; protein evolution.

PDB references: RBP-CP$_C$, 7qsp; RBP-CP$_N$, 7qsq

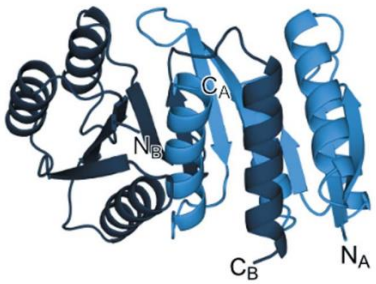
Supporting information: this article has supporting information at journals.iucr.org/d

Periplasmic binding proteins (PBPs) are a class of proteins that participate in the cellular transport of various ligands. They have been used as model systems to study mechanisms in protein evolution, such as duplication, recombination and domain swapping. It has been suggested that PBPs evolved from precursors half their size. Here, the crystal structures of two permuted halves of a modern ribose-binding protein (RBP) from *Thermotoga maritima* are reported. The overexpressed proteins are well folded and show a monomer–dimer equilibrium in solution. Their crystal structures show partially noncanonical PBP-like fold type I conformations with structural deviations from modern RBPs. One of the half variants forms a dimer via segment swapping, suggesting a high degree of malleability. The structural findings on these permuted halves support the evolutionary hypothesis that PBPs arose via a duplication event of a flavodoxin-like protein and further support a domain-swapping step that might have occurred during the evolution of the PBP-like fold, a process that is necessary to generate the characteristic motion of PBPs essential to perform their functions.

1. Introduction

Understanding the emergence of modern protein structures can be addressed by investigating the mechanisms that evolution might have employed. Some of the drivers for structural diversification are genetic mechanisms, such as mutation, duplication and recombination of domain-sized or even subdomain-sized protein fragments, offering the structural complexity needed for functions to evolve (Romero-Romero *et al.*, 2021; Sikosek & Chan, 2014; Höcker, 2014; Ohta, 2000). Another mechanism expanding this repertoire is domain swapping. While domain swapping does not lead to a change in protein sequence, its influence on the structure by forming oligomers via exchange of structural elements within the topology of a protein also contributes to the emergence of functions (Bennett *et al.*, 1995). Insights into these characteristics can shed light not only on the evolutionary history of proteins but also on our understanding of the determinants of protein folding in general.

One group of proteins that have been used for this purpose are periplasmic binding proteins (PBPs). They are involved in the cellular transport of a wide variety of small molecules such as carbohydrates, amino acids, vitamins and ions (Chandravanshi *et al.*, 2021; Felder *et al.*, 1999). The structurally symmetric bilobal architecture of their fold has long been thought to originate from a duplication and fusion event of an individual lobe (Fukami-Kobayashi *et al.*, 1999; Louie, 1993). While more detailed classifications of their fold exist (Scheepers *et al.*, 2016), they can be structurally separated into PBP-like fold types I and II, with somewhat different arrangements



of secondary-structure elements. It has been proposed that type II PBPs derive from a tandem domain swap of type I PBPs, leading to exchange of the $(\beta\alpha)_5$ elements between the lobes (Fukami-Kobayashi *et al.*, 1999). Similar domain dislocation has previously been described in related protein folds such as the chemotaxis response regulator CheY (Paithankar *et al.*, 2019), the receiver domain of cytokinin receptor CRE1 (Tran *et al.*, 2021), the tryptophan synthase subunit TrpA (Michalska *et al.*, 2020) and the uroporphyrinogen III synthase (Toledo-Patiño *et al.*, 2019; Szilágyi *et al.*, 2017).

To investigate the structural flexibility of the α/β architecture found in type I PBPs, we separated and investigated the individual lobes of the ribose-binding protein from *Thermotoga maritima* (RBP; Cuneo *et al.*, 2008). An established way to stabilize and isolate structural units within a given protein fold is the use of circular permutations (Huang, Nayak *et al.*, 2011; Iwakura *et al.*, 2000; Hennecke *et al.*, 1999). Following this approach, two protein variants that structurally represent each lobe of RBP were created and characterized (Fig. 1). We successfully obtained crystal structures of both the N-terminal lobe (RBP-CP_N) and the C-terminal lobe (RBP-CP_C), observing a non-native swapping of elements in RBP-CP_N. Our experiments also indicate dimerization of this lobe in solution, with the crystal structure showing a rearrangement reminiscent of the antiparallel β-sheet observed in type II PBPs. The observed structural malleability and the propensity to rearrange secondary-structural elements furthermore suggest a possible mechanism for transition from the type I PBP-like fold to type II via domain dislocation.

2. Materials and methods

2.1. Construct designs with *Rosetta*

The *RosettaRemodel* protocol included in the *Rosetta* suite (release 2018.19; Huang, Ban *et al.*, 2011) was used to sample possible loop conformations to connect the secondary-structure elements of the RBP lobes, leading to both the RBP-CP_N and RBP-CP_C sequences. The unliganded structure of *T. maritima* RBP (PDB entry 2fn9; Cuneo *et al.*, 2008), trimmed to include only the residues of the respective lobe, was used as a template. The new termini for the permuted

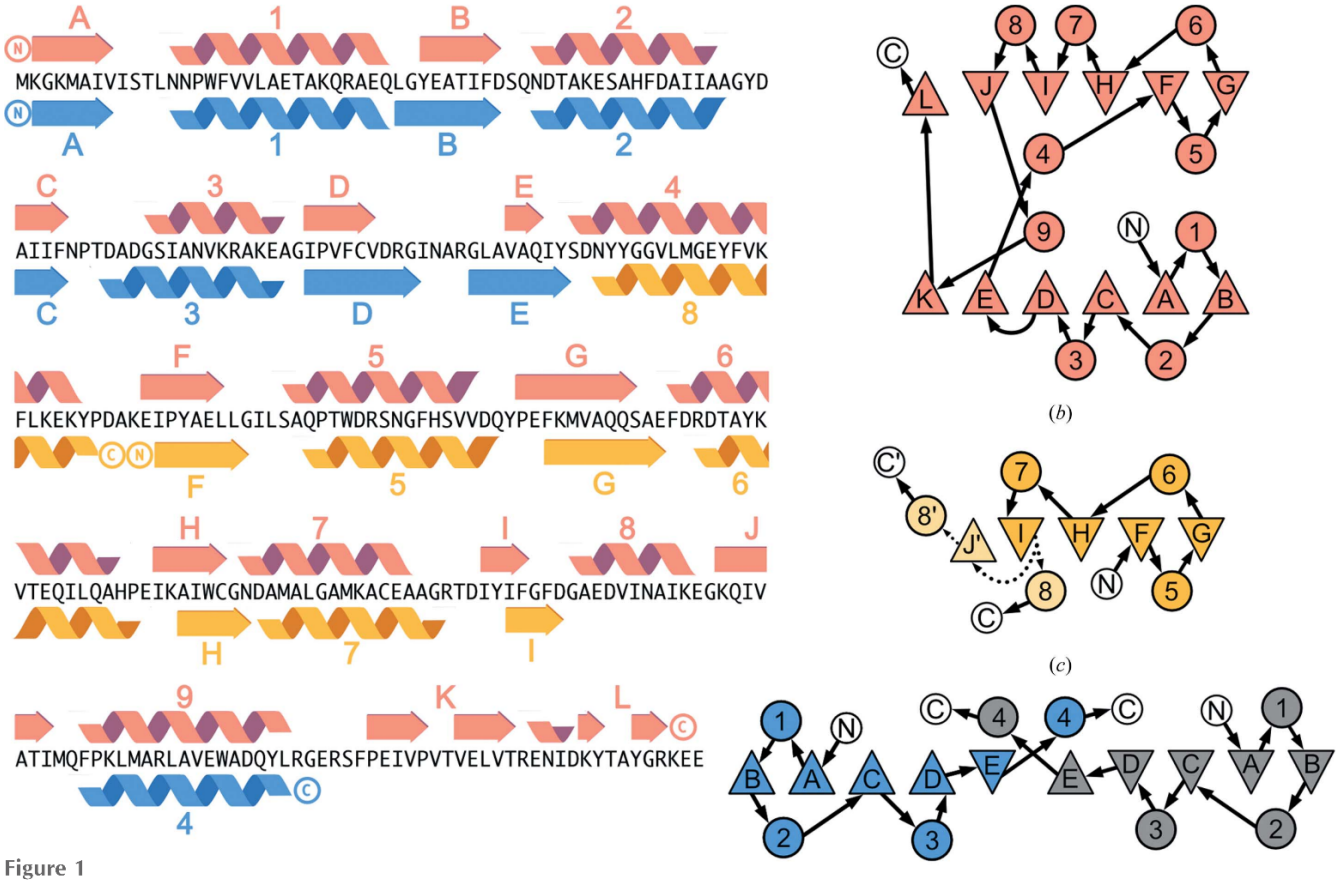


Figure 1
Secondary structure and topology of RBP and its permuted halves. (*a*) Secondary-structure alignment with the amino-acid sequence of RBP. Secondary-structure annotations derived from *PDBsum* (Laskowski *et al.*, 2018) are colored salmon for RBP, blue for RBP-CP_N and yellow for RBP-CP_C. β-Sheets are sequentially labeled with letters in the order of the sequence and α-helices are labeled with numbers. These labels correspond to the topology representation (*b*, *c*, *d*) adapted from Fukami-Kobayashi *et al.* (1999), where β-sheets are depicted as triangles and α-helices as circles. The arrangement of the secondary-structure elements reflects their three-dimensional order for RBP (*b*), RBP-CP_C (*c*) and RBP-CP_N (*d*). The N- and C-termini are labeled N and C, respectively, and the connections between the secondary-structure elements are shown as arrows. The connections of the two possible configurations of β-strand I, either to α-helix 8 or β-strand J′, in RBP-CP_C are shown as dotted arrows as these stretches are not resolved in the crystal structure.

constructs were introduced at positions 1 and 263 for RBP-CP_N, with a loop inserted between positions 105 and 244 (strand E and helix 9; Fig. 1*a*). For RBP-CP_C the N-terminus was shifted to residue 128, and a loop was inserted to connect residue 243 to the new C-terminal stretch from 106 to 127 (strand D and helix 4; Fig. 1*a*). Flexibility of the input model was allowed for one additional residue on each side of the gap during loop closure. 1000 models of three- and four-residue loops were generated using parallelized processing with Open MPI and procedural seed generation. The top ten scoring models were relaxed using the *relax* algorithm provided in this version of *Rosetta*, and the total and per-residue scoring functions were used. The sequences of the best scoring models for both RBP-CP_N and RBP-CP_C were used as final constructs (Table 1). The per-residue energies of the relaxed models were compared with the unrelaxed crystal structure of RBP and the obtained crystal structures of RBP-CP_N and RBP-CP_C using the *score_jd2* application in the same version of *Rosetta*.

Table 1
Sequences of full-length RBP and the permuted RBP halves.

The M142A mutation in RBP and the residues inserted based on *Rosetta* modeling in RBP-CP_N and RBP-CP_C are highlighted in bold.

Name	Sequence
RBP	MKGKMAIVISTLNNPWFVVLAETAKQRAEQLGYEATIFDSQND TAKESAHFDAIIAAGYDAIIFNPTDADGSIANVKRAKEAGI PVFCVDRGINARGLAVAQIYSDNYYGGVLAGEYFVKFLKEK YPDAKEIPYAELLGILS**A**QPTWDRSNGFHSVVDQYPEFKMV AQQSAEFDRDTAYKVTEQILQAHPEIKAIWCGNDAMALGAM KACEAAGRTDIYIFGFDGAEDVINAIKEGKQIVATIMQFPK LMARLAVEWADQYLRGERSFPEIVPVTVELVTRENIDKYTA YGRKLEHHHHHH
RBP-CP_N	MKGKMAIVISTLNNPWFVVLAETAKQRAEQLGYEATIFDSQND TAKESAHFDAIIAAGYDAIIFNPTDADGSIANVKRAKEAGI PVFCVDRGINARGLAVAQIYSD**TST**QFPKLMARLAVEWADQ YLRGGHHHHHH
RBP-CP_C	MKEIPYAELLGILSAQPTWDRSNGFHSVVDQYPEFKMVAQQSA EFDRDTAYKVTEQILQAHPEIKAIWCGNDAMALGAMKACEA AGRTDIYIFGFDGAEDVINAIKEGKQIVATIM**VGHNH**NYYG GVLAGEYFVKFLKEKYPDGGHHHHHH

2.2. Cloning and protein purification

The gene fragments for full-length RBP as well as RBP-CP_C were subcloned into empty linearized pET-21b(+) using NdeI/XhoI restriction sites. To prevent translation of the truncated sequence in wild-type RBP, an M142A mutation (Cuneo *et al.*, 2008) was introduced via QuikChange site-directed mutagenesis. The resulting plasmids were verified by sequencing. Gene synthesis and cloning into pET-21b(+) for RBP-CP_N were provided by Biocat. Transformant *Escherichia coli* BL21 (DE3) cells were grown in Terrific broth medium (TB) at 37°C to an OD_{600} of 1.2 in the presence of 100 µg ml^{-1} ampicillin. Protein expression was induced by the addition of 1 m*M* isopropyl *β*-D-1-thiogalactopyranoside and continued for 18 h at 20°C. The cells were harvested by centrifugation (5000*g*, 15 min), resuspended and lysed by sonication. To remove cell debris, the suspension was centrifuged again (40 000*g*, 1 h) and the supernatant was filtered through a 0.22 µm filter prior to immobilized metal ion chromatography (IMAC).

IMAC was performed on a Cytiva HisTrap 5 ml column previously equilibrated with buffer (20 m*M* MOPS, 500 m*M* NaCl, 10 m*M* imidazole pH 7.8). Elution was performed with a 40% step of elution buffer (20 m*M* MOPS, 500 m*M* NaCl, 600 m*M* imidazole pH 7.8). Fractions containing the protein of interest were pooled and concentrated for the size-exclusion chromatography step. Size-exclusion chromatography was performed on a Cytiva Superdex 26/600 75 pg column with isocratic elution of buffer (20 m*M* Tris–HCl, 300 m*M* NaCl pH 7.8). Fractions containing protein were analyzed by SDS–PAGE and those containing the proteins of interest were pooled, flash-frozen in liquid nitrogen and stored at −20°C until further analysis.

2.3. Crystallization

Initial crystallization screens were set up using a Phoenix pipetting robot (Art Robbins Instruments) with commercially available sparse-matrix screens (Qiagen; JCSG Core I–IV Suites and The PEGs Suite and PEGs Suite II) in 96-well sitting-drop plates (3-drop Intelli-Plates, Art Robbins Instruments). Droplets were pipetted in 1:1, 1:2 and 2:1 ratios of protein:reservoir solution with a protein concentration of 30 mg ml^{-1} and were incubated at 293 K. Initial crystals of RBP-CP_N appeared after 35 days in the following condition: 30% PEG 4000, 0.2 *M* lithium sulfate, 0.1 *M* Tris–HCl pH 8.5 (JCSG Core IV Suite) in the 1:1 ratio droplet. Subsequent optimization with Additive Screen (Hampton Research) yielded well diffracting cuboid-shaped crystals in the presence of the abovementioned initial hit solution supplemented with 4% 2,2,2-trifluoroethanol. Further cryoprotection was not needed.

RBP-CP_C was crystallized in the same fashion with a protein concentration of 15 mg ml^{-1}. Diffracting cuboid-shaped crystals were found after one month in 0.2 *M* magnesium acetate, 20% PEG 3350 (The PEGs Suite) in the 1:2 ratio droplet. Cryoprotection was ensured by transferring the crystal to 20% PEG 3000, 20% ethylene glycol, 0.2 *M* KNO_3.

2.4. X-ray data collection, structure determination and model building

Crystals were manually mounted using cryo-loops on SPINE standard bases and were flash-cooled after cryoprotection if needed. Diffraction data were collected on BL14.1 at the BESSY II electron-storage ring operated by the Helmholtz-Zentrum Berlin (Mueller *et al.*, 2015). Measurements were performed at 100 K in single-wavelength mode at 0.9184 Å with a Dectris PILATUS 6M detector in fine-slicing mode (0.1° wedges) using the *MXCuBE* beamline-control software (Gabadinho *et al.*, 2010). Data were processed with *XDSAPP*2 (Sparta *et al.*, 2016) employing *XDS* (Kabsch, 2010). Data quality was assessed by applying *phenix.xtriage* (Zwart *et al.*, 2005). Resolution cutoffs were determined by applying the automated paired refinement protocol *PAIREF* (Malý *et al.*, 2020).

In both cases, phases were solved by molecular replacement using the respective lobe of RBP (PDB entry 2fn9) as a search model with *Phaser* (McCoy *et al.*, 2007). The resulting models

were manually rebuilt with *Coot* (Emsley *et al.*, 2010) and refined with *phenix.refine* (Afonine *et al.*, 2018) in an iterative manner. Coordinates and structure factors were validated and deposited in the PDB (Berman *et al.*, 2002) with accession codes 7qsq (RBP-CP_N) and 7qsp (RBP-CP_C).

2.5. Far-UV circular dichroism

Far-UV circular dichroism (CD) was measured on a Jasco J-710 spectropolarimeter equipped with a Peltier device (PTC-348 WI) to control the temperature at 20°C. Before the measurements, the protein samples were dialyzed overnight into 10 m*M* sodium phosphate pH 7.8, 50 m*M* sodium chloride. Samples were measured at a protein concentration of 10 µ*M* in a 2 mm cuvette in a wavelength range from 195 to 260 nm with a bandwidth of 1 nm. After subtraction of the buffer signal, the measured ellipticity signal was converted to mean residue molar ellipticity ([Θ]) using $[\Theta] = \Theta/(lCN_r)$, where Θ is the ellipticity signal in millidegrees, l is the cell path in millimetres, C is the molar protein concentration and N_r is the number of amino acids per protein (Greenfield, 2006).

2.6. Intrinsic fluorescence

Intrinsic fluorescence (IF) spectra were collected on a Jasco FP-6500 spectrofluorometer. Measurements were performed at 20°C controlled with a water bath (Julabo MB). Samples were dialyzed and the concentration was set as described previously for CD measurements. The excitation wavelength was set to 280 nm and emission was measured in the range 300–500 nm with a bandwidth of 1 nm. The raw signal was corrected for protein concentration and further normalized to relative fluorescence.

2.7. Size-exclusion chromatography–multi-angle light scattering

Size-exclusion chromatography–multi-angle light scattering (SEC-MALS) measurements were performed with a miniDAWN detector and an Optilab refractometer (Wyatt Technology) coupled to an analytical size-exclusion chromatography column (Superdex 75 Increase 10/300 GL). Centrifuged samples were run on the column connected to an ÄKTApure FPLC system (GE Healthcare Life Sciences) and equilibrated with 10 m*M* sodium phosphate pH 7.8, 50 m*M* sodium chloride, 0.02% sodium azide at room temperature. Measurements were run at a constant flow rate of 0.8 ml min^{-1} at protein concentrations of 0.5, 1.0 and 5 mg ml^{-1}. The system setup was normalized and checked by measurement of a commercially available standardized BSA sample (2 mg ml^{-1}; Pierce, catalogue No. 23209) before and after each series of measurements. Weight-averaged molar-mass determination was performed using the Zimm equation with the differential refractive-index signal as a source for the concentration calculations (the refractive-index increment dn/dc was set to 0.185). Analysis of the experiments was performed using the *ASTRA* version 7.3.2 software suite (Wyatt Technology).

2.8. Differential scanning calorimetry

Differential scanning calorimetry (DSC) endotherms were collected using a MicroCal PEAQ-DSC instrument (Malvern Panalytical) with protein concentrations of 0.5, 1.0 and 5 mg ml^{-1}, a temperature range of 10–130°C and a scan rate of 1.5°C min^{-1}. All samples were prepared after exhaustive dialysis in 10 m*M* sodium phosphate pH 7.8, 50 m*M* sodium chloride. After proper instrument equilibration with at least two buffer–buffer scans, physical and chemical baselines were subtracted from protein–buffer scans and the data were normalized by protein concentration. *Origin* version 9.0 (OriginLab Corporation) was used for data analysis.

3. Results and discussion

3.1. Design of RBP-CP_N and RBP-CP_C

To assess how the individual lobes of a PBP-like fold behave, we chose the ribose-binding protein from *T. maritima* (RBP). Due to its thermophilic nature, it was considered to be a robust model system that could more readily tolerate this manipulation. In addition, it has previously been reported that this protein is expressed as a 21 kDa truncation (Cuneo *et al.*, 2008), suggesting that at least some elements of this protein may exist in isolation. To isolate the two lobes of RBP, the elements that make up the individual two halves were linked together via an artificial loop (Table 1). The resulting constructs RBP-CP_N (N-terminal lobe) and RBP-CP_C (C-terminal lobe) represent the two symmetric lobes of the PBP-like fold (Figs. 1*a* and 1*b*). The specific intersections were determined by structural alignment of the crystal structure of RBP from *T. maritima* in the absence of its ligand ribose (PDB entry 2fn9). RBP-CP_N was designed to consist only of the $\beta_{A–E}\alpha_{1–4}$ elements, which are directly linked to α_9. Similarly, RBP-CP_C consists of the elements $\beta_{F–J}\alpha_{5–8}$ connected to α_4 of RBP by permutation (Fig. 1*a*). To be consistent with the structure of the theoretical evolutionary precursor before duplication, the additional secondary-structural elements at the C-terminus of RBP ($\beta_{K–L}$) responsible for the second crossover between the two lobes were removed.

We obtained computational models of each lobe with comparable total and per-residue energies to the trimmed input structures of full-length RBP. Comparison of the scores obtained from the *Rosetta* energy function of native RBP and the models show similar energies for all structures (Figs. 2*a* and 2*b*). The similarity of the per-residue energy of RBP to the corresponding values for the models indicates that at least energetically, the added loop residues are suitable. The per-residue energies further show a similar distribution. For most of the sequence of RBP-CP_N, the residue energies of the crystal structure are comparable to those of the model. Only the residues of the inserted loop (blue bracket in Fig. 2*a*) score lower in the crystal structure compared to the computational model. However, the entire stretch after the inserted residues displays a higher energy (in Rosetta energy units; REU) than in the model. This is similarly reflected in both the structural rearrangement of the secondary-structure elements (Figs. 1*d*

and 2*c*) and the per-residue r.m.s.d. in RBP-CP$_N$ (Fig. 2*e*). The observation is consistent with the dimerization interface being facilitated via swapping of the α_4 element and disruption of the expected conformation at the C-terminus. While the deviation in r.m.s.d. for RBP-CP$_N$ would imply a disturbance of per-residue energies in the C-terminal stretch (Fig. 2*e*), the segment swap seems to compensate for it in canonical topology.

In contrast, a comparison of the scores of the RBP-CP$_C$ model and its resulting crystal structure shows similar energies for all resolved residues (Fig. 2*b*). The per-residue energies of the designed loop are also comparable, even though their conformation in the crystal differs significantly from the model (yellow bracket in Fig. 2*b*). Apart from the residues around the stretch of missing density (Asp96–Met116), the predicted structure corresponds well to the obtained crystal structure (Fig. 2*d*) and the per-residue r.m.s.d. values also indicate good agreement (Fig. 2*f*).

3.2. Both lobes are stable proteins with a tendency to form dimers

RBP-CP$_N$ and RBP-CP$_C$ could be expressed recombinantly in high yields in *E. coli* and purified to homogeneity. Far-UV CD spectra of both RBP-CP$_N$ and RBP-CP$_C$ show typical characteristics of a protein with an α/β-like structure and are comparable to that of full-length RBP (Fig. 3*a*). In addition, an initial hint about the correct formation of the tertiary structure in solution was obtained from the intrinsic fluorescence spectra. The emission maximum at 335 nm for both proteins as well as for RBP indicates that the aromatic residues are in a hydrophobic core and are buried from solvent, confirming that all proteins adopt a comparable compact structure (Fig. 3*b*). Another indication that the constructs appear to fold stably is the determination of thermal stability by differential scanning calorimetry (DSC). The DSC endotherms obtained for both RBP-CP$_N$ and RBP-CP$_C$ show a single and highly cooperative transition (Fig. 3*c*). The thermal unfolding appears to be irreversible, as no transition is observed upon cooling and the measurement of a second heating cycle. The permuted constructs show a lower thermostability than full-length RBP, with T_m values of 76.1 $\pm$ 0.4°C for RBP-CP$_C$ and 97.9 $\pm$ 0.9°C for RBP-CP$_N$, in contrast to 108°C for RBP (Cuneo *et al.*, 2008). There also appears to be a small dependence on protein concentration, with a shift to higher transition temperatures at higher protein concentrations (Fig. 3*c*).

Since the architecture of PBPs is likely to have originated from an ancestral dimer with the canonical binding site between the lobes, the question arises whether both variants can adopt a similar conformation. To investigate this, the oligomeric state of the proteins was determined in solution using SEC-MALS measurements (Fig. 3*d*). In the concentration range 0.5–5 mg ml^{-1}, the determined molecular weight (MW) of RBP-CP$_N$ is approximately 27.5 kDa. This

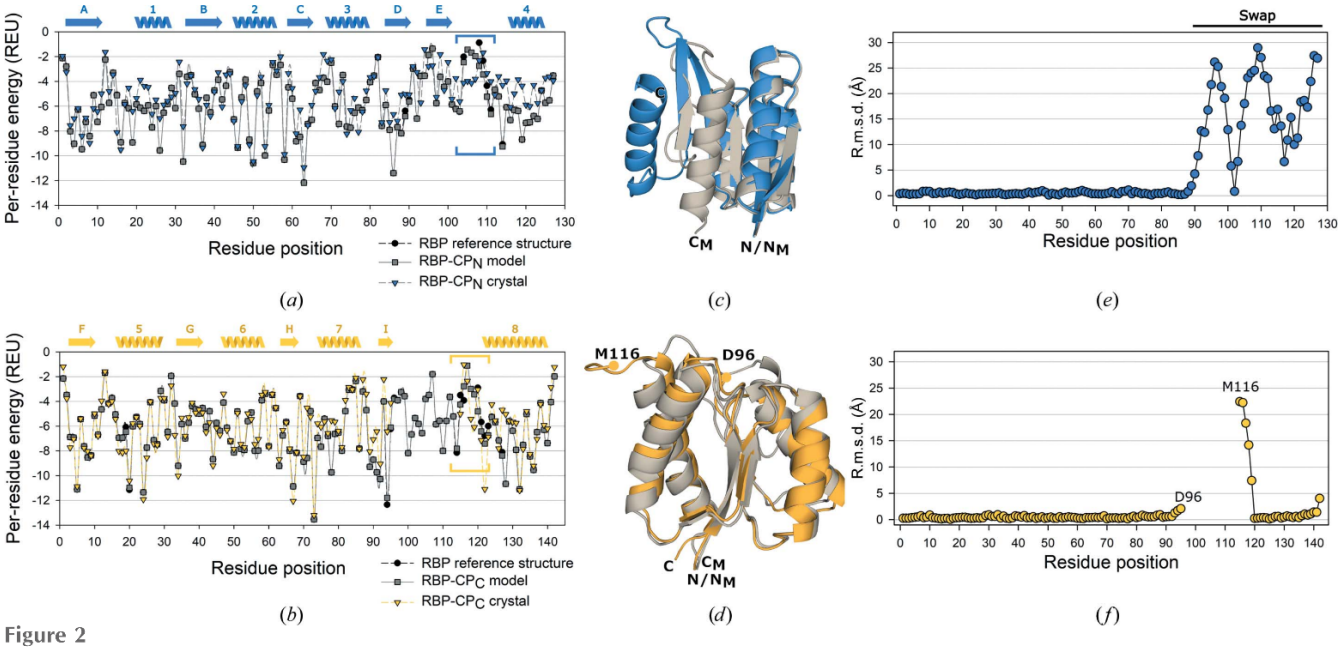


Figure 2
Per-residue *Rosetta* energy terms and comparison of the per-residue r.m.s.d. of the models to the crystal structure. (*a*, *b*) Energies in Rosetta energy units (REU) for each residue position of the template RBP structure (black, circles, dashed line), the model of RBP-CP$_N$ or RBP-CP$_C$ (gray, squares, solid line) and the respective crystal structures (blue for RBP-CP$_N$ and yellow for RBP-CP$_C$, triangles, dashed lines). Sites where loop residues were introduced are highlighted by colored brackets for each protein. Secondary-structural elements as observed in the crystal are shown and are labeled as in Fig. 1(*a*). (*c*, *d*) Superposition of the computational models (gray) and the corresponding crystal structures of RBP-CP$_N$ (blue) and RBP-CP$_C$ (yellow). The borders of the area of missing density in RBP-CP$_C$ are labeled D96 and M116. (*e*, *f*) Per-residue r.m.s.d. (based on C$^\alpha$ atoms) of the obtained crystal structures of RBP-CP$_N$ (blue) and RBP-CP$_C$ (yellow) compared with their models. The representation and alignment were obtained using *PyMOL* 2.5.0 (Schrödinger) and the `align` command with `cycles=0`, considering only C$^\alpha$ atoms of chains *A* and transferring per-residue values with the `rmsd_b` script (http://pldserver1.biochem.queensu.ca/~rlc/work/pymol/rmsd_b.py).

Table 2
Molecular-weight determination with SEC-MALS.

Sample (concentration)	Expected MW (kDa)	Experimental MW (kDa)	Uncertainty (%)
RBP-CP_N (0.5 mg ml^{-1})	14.9	26.8	0.8
RBP-CP_N (1.0 mg ml^{-1})		27.2	0.5
RBP-CP_N (5.0 mg ml^{-1})		28.5	0.3
RBP-CP_C (0.5 mg ml^{-1})	16.7	18.0	1.0
RBP-CP_C (1.0 mg ml^{-1})		18.7	0.7
RBP-CP_C (5.0 mg ml^{-1})		22.4	0.4

corresponds to a dimeric conformation, as it is about double the expected monomeric MW of 14.9 kDa. The shift from lower molecular weight at lower concentrations to higher molecular weight at higher concentrations indicates that the monomer–dimer equilibrium is dynamic and concentration-dependent. A similar pattern is observed for RBP-CP_C. While the protein appears to be monomeric at low concentrations (0.5 mg ml^{-1}), the MW shifts to 18.7 kDa at 1 mg ml^{-1} and to 22.4 kDa at 5 mg ml^{-1}. This would correspond to a dynamic shift from a monomer (theoretical MW of 16.7 kDa) to a dimer (Table 2). These results are in agreement with the concentration-dependent thermostability observed in DSC measurements. Together, they explain the shift to higher temperatures during thermal unfolding, with possible stabilization of the overall fold by forming a defined dimer interface.

3.3. The structures of both RBP-CP_N and RBP-CP_C differ from their native counterparts

The PBP-like type I canonical fold consists of two lobes with a continuous, parallel β-sheet with five strands in the order 21345 plus an additional, noncontinuous β_6 strand flanked by alternating α-helices on each side and one crossover between each lobe (Figs. 1*a* and 1*b*). In contrast to the expected single-lobed architecture, the crystal structures obtained for RBP-CP_N and RBP-CP_C deviate from the structure of full-length RBP.

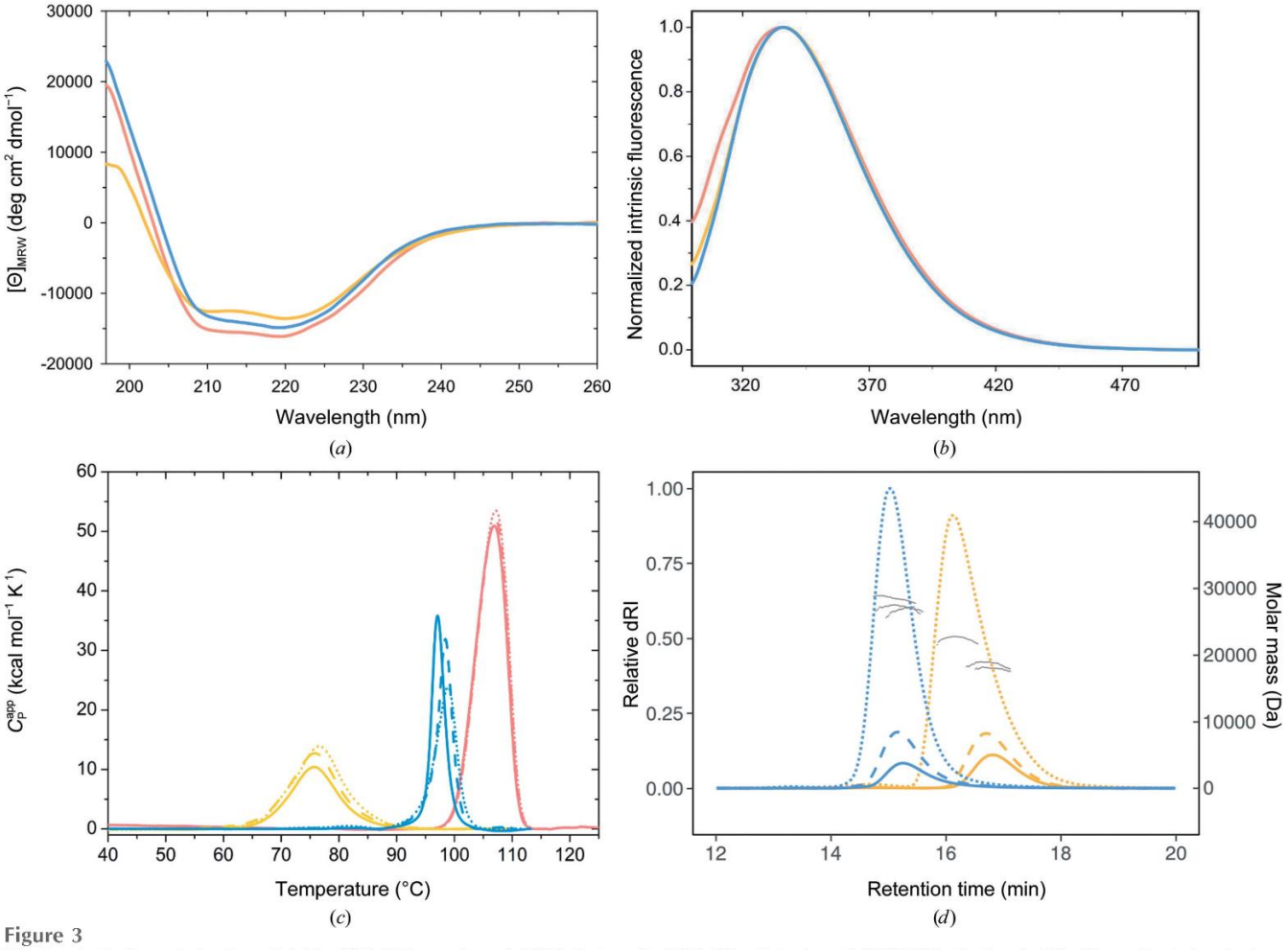


Figure 3
Biochemical characterization. (*a*) Far-UV CD spectra of RBP (salmon), RBP-CP_N (blue) and RBP-CP_C (yellow). (*b*) Normalized tryptophan fluorescence at a 280 nm excitation wavelength of RBP (salmon), RBP-CP_N (blue) and RBP-CP_C (yellow). (*c*) DSC endotherms of RBP (salmon), RBP-CP_N (blue) and RBP-CP_C (yellow); sample concentrations of 0.5, 1 and 5 mg ml^{-1} are shown as solid, dashed and dotted lines, respectively. (*d*) SEC-MALS analysis of RBP-CP_N (blue) and RBP-CP_C (yellow) at different concentrations. The elution profile is plotted as the relative differential refractive index against the retention time. Sample concentrations of 0.5, 1 and 5 mg ml^{-1} are shown as solid, dashed and dotted lines, respectively. Molar-mass determinations for peak regions are plotted as gray dots.

Table 3
Crystallographic data and refinement statistics.

	RBP-CP$_N$	RBP-CP$_C$
PDB code	7qsq	7qsp
Wavelength (Å)	0.9184	0.9184
Resolution range (Å)	48.96–1.79 (1.86–1.79)	39.76–1.36 (1.40–1.36)
Space group	$P2_1$	$P2_12_12_1$
a, b, c (Å)	55.37, 62.77, 76.26	41.69, 41.97, 132.20
α, β, γ (°)	90, 102.1, 90	90, 90, 90
Total reflections	176181 (15604)	533879 (48154)
Unique reflections	47556 (4346)	50883 (4875)
Multiplicity	3.7 (3.6)	10.5 (9.9)
Completeness (%)	97.8 (85.5)	99.0 (96.7)
Mean $I/\sigma(I)$	8.58 (0.76)	13.93 (1.00)
Wilson B factor (Å^2)	32.6	18.8
No. of molecules in asymmetric unit	4	2
Matthews coefficient (Å^3 Da^{-1})	2.14	1.72
R_{merge}	0.080 (1.324)	0.081 (1.907)
R_{meas}	0.094 (1.548)	0.085 (2.008)
$R_{p.i.m.}$	0.047 (0.788)	0.026 (0.616)
$CC_{1/2}$	0.997 (0.413)	0.999 (0.322)
CC*	0.999 (0.765)	1.000 (0.698)
Reflections used in refinement	47294 (4102)	50883 (4875)
Reflections used for R_{free}	2088 (181)	2100 (201)
R_{work}	0.191 (0.370)	0.171 (0.353)
R_{free}	0.239 (0.396)	0.210 (0.380)
CC_{work}	0.963 (0.685)	0.962 (0.605)
CC_{free}	0.952 (0.532)	0.938 (0.554)
No. of non-H atoms		
Total	4446	2308
Macromolecules	4063	2073
Solvent	315	199
No. of protein residues	510	248
R.m.s.d., bond lengths (Å)	0.003	0.012
R.m.s.d., bond angles (°)	0.57	1.23
Ramachandran favored (%)	98.4	99.2
Ramachandran allowed (%)	1.4	0.8
Ramachandran outliers (%)	0.2	0.0
Rotamer outliers (%)	1.4	1.4
Clashscore	5.16	4.82
Average B factor (Å^2)		
Overall	40.0	25.9
Macromolecules	39.2	24.6
Solvent	46.5	35.9
No. of TLS groups	4	2

RBP-CP$_C$ crystallized in the orthorhombic space group $P2_12_12_1$, with two chains of the protein in the asymmetric unit, and was refined to a resolution of 1.36 Å (Table 3). While the N-terminal $(\alpha\beta)_4$ elements in both chains are nearly identical to the core of the corresponding part in full-length RBP, the remaining elements differ from the canonical topology (Figs. 1*b* and 1*c*). While the core structure of $\alpha_{5–7}$ and $\beta_{F–I}$ in RBP-CP$_C$ is comparable to that of RBP, the following β_J strand and the synthetic loop are not resolved in the crystal structure (Fig. 4*a*). However, the connecting α_8 helix on the other side of this gap in the structure can unambiguously be seen (Fig. 1*c*). It remains unclear whether the inserted loop or the energetical frustration of missing elements on this terminal side of the protein interferes with the proper formation of β_J, or whether a preferential but unobserved swap of elements with an adjacent protein molecule results in the lack of density in this protein region (Fig. 4*e*). An alternative explanation could be the formation of an interface between two crystallographic dimers, as indicated by an analysis with the *PISA* server (Krissinel & Henrick, 2007). In this case, the C-terminal α_8 would not originate from the same chain of the asymmetric unit but from its corresponding symmetry mate. The resulting extended arrangement is facilitated by an interaction of the β_1 strand and the residue stretch 116′–120′ (Fig. 5*a*). This extension is similar to a continuation of the sheet via the antiparallel addition of a short, single stretch resembling a strand, with the residues of the designed loop (Val117–His121) participating in the interaction (Fig. 1*c*). With the α_4 helix originating from the adjacent symmetry mate, it is also possible that there is a mixed population of both conformations, with the helix serving as a common structural anchor point. This could also explain the lack of density in the connecting area. A similar shuffling of elements can be observed with less ambiguity in the crystal structure of RBP-CP$_N$ (Fig. 5*b*). This possible interaction could also explain the concentration-dependent oligomerization observed in the SEC-MALS measurements (Fig. 3*d*). The central β-sheet as well as all α-helices appear to be well ordered, except for the loops close to the unresolved region and the termini. The r.m.s.d. of 0.5 Å over 135 C$^\alpha$ atoms of the resolved residues, however, indicates a high similarity between RBP-CP$_C$ and the corresponding elements of full-length RBP (Fig. 4*c*).

The case is different when looking at the N-terminal lobe. The crystal structure of RBP-CP$_N$ was solved in the monoclinic space group $P2_1$ at 1.79 Å resolution. The asymmetric unit is composed of four chains, of which two pairs form a dimer via a segment swap. Unlike the interface of the two lobes in native PBPs, the dimer is located on the edge of the two central β-sheets (Fig. 4*b*). This extension of the sheet is mediated via each of the respective β_E strands. In contrast to the rest of the central β-sheet, the two β_E strands form an antiparallel stretch of the extended β-sheet. This change in direction of the C-terminal β-strand is not known to occur in PBP-like fold type I proteins, in which the central β-sheet always adopts a parallel conformation. In addition, this swap of the $\beta_D\beta_E$ elements in their parallel–antiparallel arrangement forms the interface of the dimer (Fig. 1*d*). These structural rearrangements are also reflected by the significant difference in r.m.s.d. of 5.9 Å when comparing the structure of RBP-CP$_N$ with the equivalent half of the full-length RBP (Fig. 4*d*). This unusual rearrangement of elements indicates a high tolerance of this structural motif to variations in its topology. In agreement with other structures, such as the CheY-like fold (Paithankar *et al.*, 2019), the TIM-barrel fold (Michalska *et al.*, 2020) and other related folds (Lewis *et al.*, 2000; Tran *et al.*, 2021; Szilágyi *et al.*, 2017), the isolated domains of a PBP-like type I protein show a high degree of malleability.

4. Conclusions

The obtained crystal structures of the permuted constructs of both the N- and C-terminal lobes of RBP from *T. maritima* suggest the possibility that they could have existed in isolation of the full structural context. This corresponds to the idea that modern PBPs arose from a duplication event. Based on

structural and sequence similarities, it has been proposed that this progenitor was an ancestral protein of the flavodoxin-like fold. The existence of the stable permuted halves clearly shows that the single lobe can exist on its own and can help inform on this evolutionary process.

However, the observed swapping of elements in RBP-CP$_N$ could also correspond to another event in the evolution of PBPs. It has previously been concluded that the evolution of the PBP-like fold involved domain swapping of the C-terminal helices, a step that was necessary to generate the characteristic

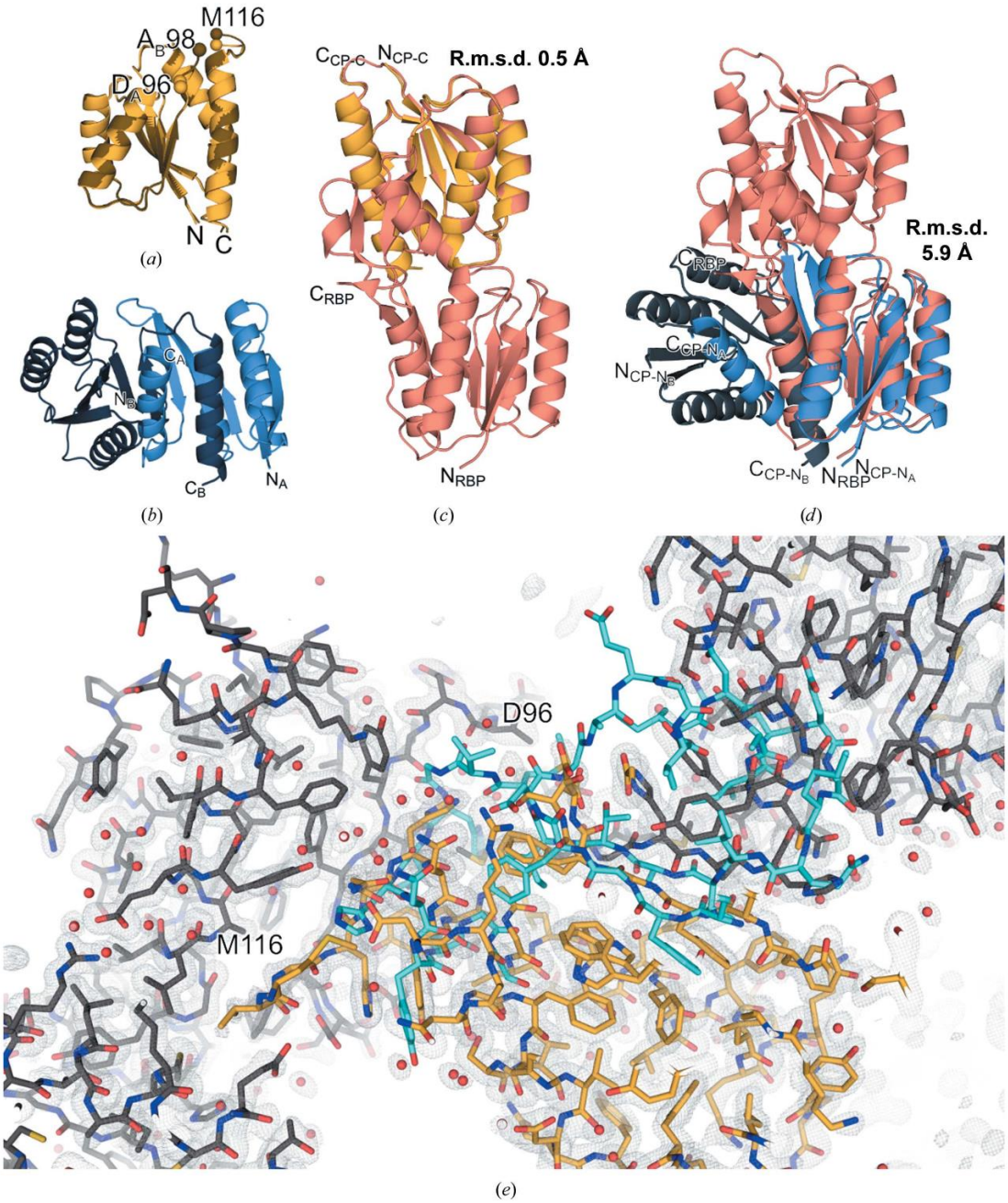


Figure 4
Comparison of the crystal structures of the individual lobes with full-length RBP. (*a*) Cartoon representation of the structural alignment of the two chains in the asymmetric unit of RBP-CP$_C$, with the edges of the unresolved region of chains *A* (Asp95–Met116) and *B* (Ala98–Met116) shown as spheres. (*b*) Cartoon representation of the crystallographic dimer of RBP-CP$_N$. (*c*, *d*) Superposition of the cartoon structures of full-length RBP with RBP-CP$_C$ (*c*) and RBP-CP$_N$ (*d*), respectively. R.m.s.d. values over all C$^\alpha$ atoms of chain *A* of each structure are provided next to each figure. (*e*) Missing density in the RBP-CP$_C$ map spanning residues Asp96–Met116. The crystal structure is shown as sticks, where chain *A* is colored yellow and symmetry mates are colored gray. A stick representation of the corresponding *Rosetta* model (residues Ile92–Gly125) is shown as an overlay in cyan. Water molecules are depicted as red spheres. A $2F_o - F_c$ map contoured at an r.m.s.d. of 1.0 is shown as gray mesh. The representation and alignment were obtained using *PyMOL* 2.3.0 (Schrödinger) and the `align` command with `cycles=0`.

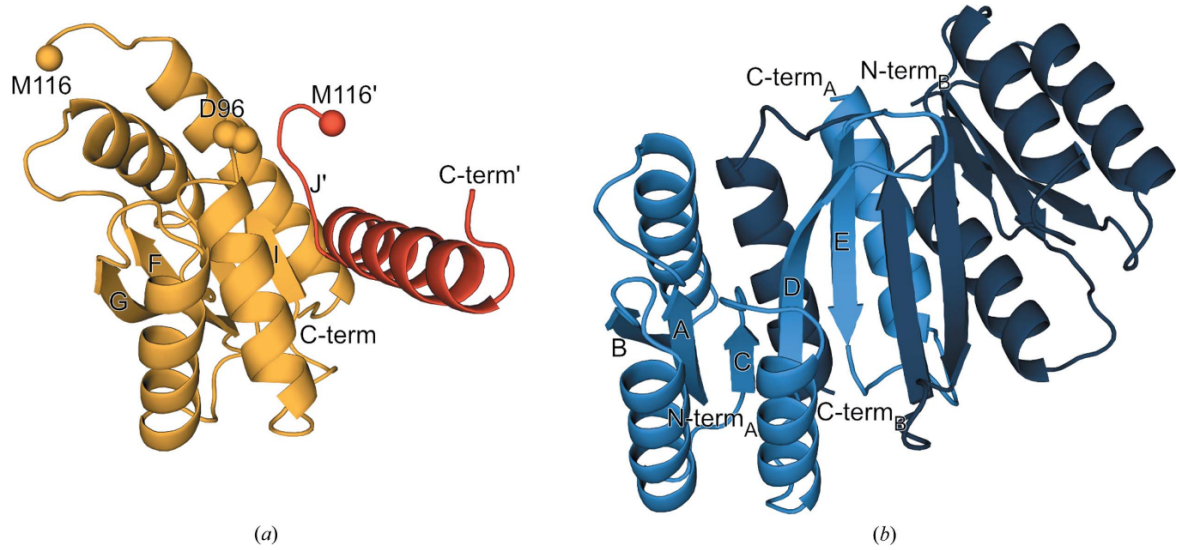


Figure 5
Possible alternative interface facilitated by a symmetry mate in the crystal structure of RBP-CP$_C$. (*a*) Cartoon representation of the interface of chain *A* and the participating elements of its symmetry mate chain *A*′ in the crystal structure of RBP-CP$_C$. The termini of the protein and the gap where the chain could not be traced are labeled for each chain. (*b*) Cartoon representation of the interface of the RBP-CP$_N$ dimer. Secondary structures are labeled according to Fig. 1.

hinge-bending motion of PBPs, with subsequent fusion of this proposed ancestral dimer (Fukami-Kobayashi *et al.*, 1999). In addition, it has been proposed that the absence of the helix between β-strands D and E and helix 8 (Fig. 1*b*) may have been a necessary step for the swapping event that led to PBPs with the type II fold. This partially explains why we observe a dimer with an unusual segment swap in RBP-CP$_N$, which lacks this helix. However, it appears that RBP-CP$_C$, which still contains this corresponding helix 8, does not reliably form a dimer. However, the alternative interface involving the chain from a symmetry mate could partially explain the behavior observed in SEC-MALS measurements. The dynamic shift to higher molecular weight species can only be observed at high protein concentrations. Interestingly, however, the antiparallel stretch of residues 117′–119′ in RBP-CP$_C$ bears a resemblance to the continuation of the central β-sheet in RBP-CP$_N$. The residues participating in the interaction with β4 are the additional residues introduced via the design. A reason for this could be the energetically frustrated surface of β4, which now lacks the corresponding β5 from RBP, that induces the switch of the designed loop into a more strand-like conformation to satisfy this hydrophobic surface.

Alternatively, a possible explanation may lie in the folding pathway of proteins with a flavodoxin-like fold. The folding mechanism of CheY, a well studied protein with a flavodoxin-like fold, suggests that there may be a universal subdomain intermediate in the folding pathway (Hills & Brooks, 2008). The N-terminal $\beta_{1-3}\alpha_{1-2}$ elements appear to initially form a central triad followed by folding of the remaining elements. The permuted RBP lobes could follow a similar path. The corresponding elements could form a folded scaffold onto which the rest of the protein folds. This substructure potentially stabilizes the protein to a point where the C-terminal elements can still adapt a structured conformation but provide sufficient flexibility for the unusual rearrangement that we have found.

The novel antiparallel stretch of the dimer-swapped β-sheets has not been observed before in proteins with the type I PBP-like fold, and the existence of this swap highlights the flexibility of this structural element. Additionally, the alleviation of the energetically frustrated hydrophobic surface achieved via the alternative interface in the structure of RBP-CP$_C$ could offer valuable insights into the mechanisms behind domain swapping in PBPs in general. More detailed sequence analysis and experiments would be required to obtain a clear picture of the transition from type I to type II PBPs. The malleability of this α/β architecture, which is also apparent in other folds (for example the Rossmann, flavodoxin and TIM-barrel-like folds), may be a reason for its frequent occurrence in modern proteins (Ferruz *et al.*, 2021).

Acknowledgements

We acknowledge the allocation of synchrotron beamtime and financial support by HZB and thank the beamline staff at BESSY for support. The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper. We thank all members of the Höcker Laboratory for their constructive suggestions to improve the research. Open access funding enabled and organized by Projekt DEAL.

Funding information

This work was supported by the European Research Council (ERC Consolidator Grant 647548 'Protein Lego' to BH), the VolkswagenStiftung (grant 94747 to BH) and by a fellowship

from the Alexander von Humboldt and Bayer Science and Education Foundations (Humboldt–Bayer Research Fellowship for Postdoctoral Researchers to SRR).

References

Afonine, P. V., Poon, B. K., Read, R. J., Sobolev, O. V., Terwilliger, T. C., Urzhumtsev, A. & Adams, P. D. (2018). *Acta Cryst.* D**74**, 531–544.

Bennett, M. J., Schlunegger, M. P. & Eisenberg, D. (1995). *Protein Sci.* **4**, 2455–2468.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 899–907.

Chandravanshi, M., Tripathi, S. K. & Kanaujia, S. P. (2021). *FEBS Lett.* **595**, 2395–2409.

Cuneo, M. J., Beese, L. S. & Hellinga, H. W. (2008). *BMC Struct. Biol.* **8**, 50.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Felder, C. B., Graul, R. C., Lee, A. Y., Merkle, H. P. & Sadee, W. (1999). *AAPS PharmSci*, **1**, E2.

Ferruz, N., Michel, F., Lobos, F., Schmidt, S. & Höcker, B. (2021). *Front. Mol. Biosci.* **8**, 715972.

Fukami-Kobayashi, K., Tateno, Y. & Nishikawa, K. (1999). *J. Mol. Biol.* **286**, 279–290.

Gabadinho, J., Beteva, A., Guijarro, M., Rey-Bakaikoa, V., Spruce, D., Bowler, M. W., Brockhauser, S., Flot, D., Gordon, E. J., Hall, D. R., Lavault, B., McCarthy, A. A., McCarthy, J., Mitchell, E., Monaco, S., Mueller-Dieckmann, C., Nurizzo, D., Ravelli, R. B. G., Thibault, X., Walsh, M. A., Leonard, G. A. & McSweeney, S. M. (2010). *J. Synchrotron Rad.* **17**, 700–707.

Greenfield, N. J. (2006). *Nat. Protoc.* **1**, 2876–2890.

Hennecke, J., Sebbel, P. & Glockshuber, R. (1999). *J. Mol. Biol.* **286**, 1197–1215.

Hills, R. D. Jr & Brooks, C. L. III (2008). *J. Mol. Biol.* **382**, 485–495.

Höcker, B. (2014). *Curr. Opin. Struct. Biol.* **27**, 56–62.

Huang, P.-S., Ban, Y.-E. A., Richter, F., Andre, I., Vernon, R., Schief, W. R. & Baker, D. (2011). *PLoS One*, **6**, e24109.

Huang, Y.-M., Nayak, S. & Bystroff, C. (2011). *Protein Sci.* **20**, 1775–1780.

Iwakura, M., Nakamura, T., Yamane, C. & Maki, K. (2000). *Nat. Struct. Biol.* **7**, 580–585.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.

Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.

Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S. & Thornton, J. M. (2018). *Protein Sci.* **27**, 129–134.

Lewis, R. J., Muchová, K., Brannigan, J. A., Barák, I., Leonard, G. & Wilkinson, A. J. (2000). *J. Mol. Biol.* **297**, 757–770.

Louie, G. V. (1993). *Curr. Opin. Struct. Biol.* **3**, 401–408.

Malý, M., Diederichs, K., Dohnálek, J. & Kolenko, P. (2020). *IUCrJ*, **7**, 681–692.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

Michalska, K., Kowiel, M., Bigelow, L., Endres, M., Gilski, M., Jaskolski, M. & Joachimiak, A. (2020). *Acta Cryst.* D**76**, 166–175.

Mueller, U., Förster, R., Hellmig, M., Huschmann, F. U., Kastner, A., Malecki, P., Pühringer, S., Röwer, M., Sparta, K., Steffien, M., Ühlein, M., Wilk, P. & Weiss, M. S. (2015). *Eur. Phys. J. Plus*, **130**, 141–151.

Ohta, T. (2000). *Philos. Trans. R. Soc. London B*, **355**, 1623–1626.

Paithankar, K. S., Enderle, M., Wirthensohn, D. C., Miller, A., Schlesner, M., Pfeiffer, F., Rittner, A., Grininger, M. & Oesterhelt, D. (2019). *Acta Cryst.* F**75**, 576–585.

Romero-Romero, S., Kordes, S., Michel, F. & Höcker, B. (2021). *Curr. Opin. Struct. Biol.* **68**, 94–104.

Scheepers, G. H., Lycklama a Nijeholt, J. A. & Poolman, B. (2016). *FEBS Lett.* **590**, 4393–4401.

Sikosek, T. & Chan, H. S. (2014). *J. R. Soc. Interface*, **11**, 20140419.

Sparta, K. M., Krug, M., Heinemann, U., Mueller, U. & Weiss, M. S. (2016). *J. Appl. Cryst.* **49**, 1085–1092.

Szilágyi, A., Györffy, D. & Závodszky, P. (2017). *Proteins*, **85**, 46–53.

Toledo-Patiño, S., Chaubey, M., Coles, M. & Höcker, B. (2019). *Biochemistry*, **58**, 4790–4793.

Tran, L. H., Urbanowicz, A., Jasiński, M., Jaskolski, M. & Ruszkowski, M. (2021). *Front. Plant Sci.* **12**, 756341.

Zwart, P. H., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *CCP4 Newsl. Protein Crystallogr.* **43**, contribution 7.

6. Paper III

Ferruz N.**, Michel F.**, Lobos F., Schmidt S., Höcker B.
Fuzzle 2.0: Ligand Binding in Natural Protein Building Blocks.
Frontiers in Molecular Biosciences. 2021, 18;8:715972

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 18 August 2021
doi: 10.3389/fmolb.2021.715972

Fuzzle 2.0: Ligand Binding in Natural Protein Building Blocks

Noelia Ferruz[1]*, *Florian Michel*[1], *Francisco Lobos*[1], *Steffen Schmidt*[2] and *Birte Höcker*[1]*

[1]Department of Biochemistry, University of Bayreuth, Bayreuth, Germany, [2]Computational Biochemistry, University of Bayreuth, Bayreuth, Germany

Modern proteins have been shown to share evolutionary relationships *via* subdomain-sized fragments. The assembly of such fragments through duplication and recombination events led to the complex structures and functions we observe today. We previously implemented a pipeline that identified more than 1,000 of these fragments that are shared by different protein folds and developed a web interface to analyze and search for them. This resource named Fuzzle helps structural and evolutionary biologists to identify and analyze conserved parts of a protein but it also provides protein engineers with building blocks for example to design proteins by fragment combination. Here, we describe a new version of this web resource that was extended to include ligand information. This addition is a significant asset to the database since now protein fragments that bind specific ligands can be identified and analyzed. Often the mode of ligand binding is conserved in proteins thereby supporting a common evolutionary origin. The same can now be explored for subdomain-sized fragments within this database. This ligand binding information can also be used in protein engineering to graft binding pockets into other protein scaffolds or to transfer functional sites *via* recombination of a specific fragment. Fuzzle 2.0 is freely available at https://fuzzle.uni-bayreuth.de/2.0.

Keywords: web server, protein evolution, protein design, protein fragment, flavodoxin-like fold, periplasmic binding protein

OPEN ACCESS

Edited by:
Brian Jiménez-García, Utrecht University, Netherlands

Reviewed by:
Alexander Goncearenco, National Center for Biotechnology Information (NLM), United States
Manon Réau, Utrecht University, Netherlands
Jorge Roel, Instituto de Biología Molecular de Barcelona (IBMB), Spain

***Correspondence:**
Birte Höcker
birte.hoecker@uni-bayreuth.de
Noelia Ferruz
noelia.ferruz-capapey@uni-bayreuth.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *27 May 2021*
Accepted: *06 August 2021*
Published: *18 August 2021*

Citation:
Ferruz N, Michel F, Lobos F, Schmidt S and Höcker B (2021) Fuzzle 2.0: Ligand Binding in Natural Protein Building Blocks. Front. Mol. Biosci. 8:715972. doi: 10.3389/fmolb.2021.715972

INTRODUCTION

A main function of proteins is the binding of molecules such as other proteins or smaller compounds. For example, the entire machinery of metabolic pathways consists of proteins that bind various substrates and catalyze diverse reactions (Schmidt and Dandekar, 2002). Despite this apparent diversity proteins were often reused in the course of evolution and their reactions adapted to perform different functions. In fact, todays diverse set of proteins and their associated functions are the product of mutation, recombination and duplication events (Horowitz, 1945; Jensen, 1976; Ohta, 2000; Sikosek and Chan, 2014).

For a long time, protein domains have been considered as the evolutionary unit, being structurally discrete and independently folding. However, the analysis of the known sequence and structure space in recent years led to a renewed insight on an old concept: Modern proteins might have arisen from a set of primordial peptides to increasingly larger subdomain-sized fragments (Alva and Lupas, 2018; Romero-Romero et al., 2021). Based on sequence and structural similarities it is possible to infer likely evolutionary relationships of proteins, even of different folds (Farías-Rico et al., 2014). The examples provided by Farías-Rico et al. and Alva et al. show how nature used these ready-made pieces in the evolution of modern protein diversity.

A number of studies have now identified several subdomain-sized fragments as common evolutionary units (Alva et al., 2015; Nepomnyachiy et al., 2017; Ferruz et al., 2020). The database of subdomain-sized fragments that we developed previously is accessible *via* a web interface to allow

individual analysis (Ferruz et al., 2020). These conserved fragments often participate in ligand binding, including nucleotides, nucleotide-derived cofactors, or metal ions (Bharat et al., 2008; Laurino et al., 2016; Goncearenco and Berezovsky, 2015; Romero Romero et al., 2018; Longo et al., 2020; Narunsky et al., 2020). This clearly indicates a key role of ligand interactions in the evolution of these ancestral building blocks.

To include this important aspect, we have updated Fuzzle to allow systematic searches for ligands and to enable a better understanding of the evolution of protein fragments. Fuzzle 2.0 enables the analysis of non-covalent interactions of protein-ligand complexes. Additionally, it now also allows searching for homologous fragments that nature has reused as building blocks that bind the same ligand. Here, we demonstrate its new capabilities using as an example a periplasmic binding protein (PBP). We show how PBPs contain a conserved fragment that is associated with several ligands and we highlight its homologous relationships to several other superfamilies. This conserved protein building block is examined from an evolutionary as well as a protein engineering perspective.

MATERIALS AND METHODS

Database

The Fuzzle database uses SCOPe (Fox et al., 2014) to identify protein domains. SCOPe is a hierarchical database that sorts domains into folds, superfamilies and families. We first updated Fuzzle to include SCOPe release 2.07. Common sub-domain fragments were identified as previously described (Ferruz et al., 2020). In particular, we created hidden Markov model profiles for each domain in SCOP95 2.07 using the HH-suite (Söding, 2005). These domains were compared all-against-all using HHsearch and then structurally superimposed using TM-align (Zhang and Skolnick, 2005). TM-align calculates the RMSD based on Cα-atoms. The data is stored in the database as 'SCOPe 2.07 PSI'. We then filtered hits (pairs of domains that have a fragment in common) from different folds, with an RMSD <3 Å, HHsearch probability over 70%, length between 10 and 200 amino acids and TM-score > 0.3. Hits were allowed to have sequence alignments at most 25% longer than the structural alignments. Since SCOPe lacks coordinates of bound ligands, we retrieved the coordinates from the original PDB entries using a 4 Å distance cutoff for any heavy atom. To stay consistent with the PDB definition, all 'HETATM' entries were considered as ligands, including modified residues. We added the corresponding ligands' coordinates. In cases where a ligand is bound in between multiple domains, it will appear with all domains where it shows an interaction based on the cutoff.

Website

The web interface contains several updates from its predecessor version. It is now possible to search for ligands in two ways: either by its PDB (three-letter) code (e.g., ATP for adenosine-5′-triphosphate) or by its SMILES (Weininger, 1988). SMILES searches in Fuzzle 2.0 not only find ligands that are identical, but users can also search ligands that are more than 70% similar. Similarity searches use topological RDKit fingerprints (default parameters: minimum path size: 1 bond - maximum path size: 7 bonds - fingerprint size: 2048 bits - number of bits set per hash: 2 - minimum fingerprint size: 64 bits - target on-bit density 0.0) with Tanimoto similarity coefficient (Godden et al., 2000). Moreover, SMILES searches allow to identify sub- or superstructures of a compound (e.g., adenosine and inorganic phosphate as substructures of ATP).

The database has now been extended to include additional information about ligands and fragments. For example, the fragment analysis page now contains a table that includes the statistics of the fragment: A representative domain that contains each fragment (selected as the domain with most network connections to other domains that also contain that fragment, such as domain 'd1jw9b_' for fragment 1: https://fuzzle.uni-bayreuth.de/2.0/fragments/network/fragment/1), the number of domains that contain the fragment, the average fragment length, involved folds, and the ligands bound to the fragment. In a detailed view it is possible to visualize protein-ligand interactions in the context of fragments using the NGL Viewer (Rose et al., 2018). To analyze the interactions, one can toggle different interaction types, compute distances, and show surface representations. The relationship tables between all SCOPe categories have been updated to reflect the ligand information. Tables and networks containing this updated information can be downloaded as CSV or JSON files. Superpositions of fragments are available as PyMOL sessions. Additionally, a fragment search was implemented to allow finding fragments depending on their ligands, SCOPe category or length. The web frontend was altered to reflect these changes. To this end, we use Django (version 1.11), PostgreSQL, and JavaScript. The style of the web site relies on the Bootstrap framework (version 4.0). Other software technologies used in Fuzzle 2.0 include JQuery (jquery.com), graph_tool (graph-tool.skewed.de), Datatables (datatables.net), and D3js (d3js.org) to visualize the data.

Analysis of Ligands Binding the PBP-like Fragment

Ligands that are commonly known to be additives to crystallization screens or other experiments were excluded from our ligand-binding analysis to the PBP fragment. These additives are listed in BioLiP (Yang et al., 2013) and in this case correspond to: ACM, ACT, ACY, CIT, CL, EDO, FMT, GOL, MPD, MPO, MSE, NA, PEG, PO4, SO4, and TRS. CA and MSO were also removed from the set. Sequence alignments in the **Supplementary Material** were retrieved from each of the pairwise alignments to d2fn9a_ and grouped by superfamily. The webpage allows to filter out these crystallization artifacts and post-translational modifications with toggle buttons (e.g: https://fuzzle.uni-bayreuth.de/2.0/fragments/table/).

RESULTS

New Fuzzle Features

The original Fuzzle database already contained a large number of conserved fragments that are shared between folds thereby illustrating a remarkable connectivity of the protein universe.

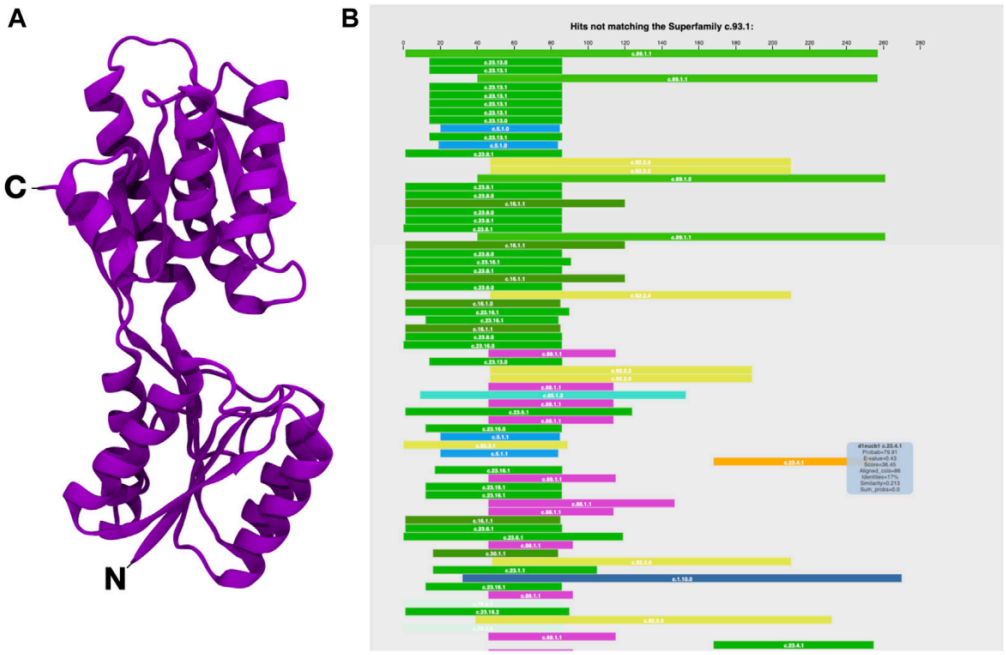


FIGURE 1 | The ribose binding protein from *Thermotoga maritima* (tmRBP, d2fn9a_). **(A)** Structure (PDB 2FN9) and **(B)** HHsearch overview (https://fuzzle.uni-bayreuth.de/2.0/hh/hhs-raw/d2fn9a_/). The query sequence (d2fn9a_) is taken as reference for the alignment of the proteins found (residues 1–280). Each bar represents a hit that is related with the query domain but does not belong to the same superfamily. The position of the bar matches the location of the hit related to the query's sequence. Colors differentiate superfamilies. Extra information on the hits can be found by mousing over. Note that in this summary diagram hits that were found to other proteins of the same superfamily are not shown.

The inclusion of the SCOPe 2.07 database increased the size of domains and thereby the number of pairwise hits (**Supplementary Table S1**). As with the previous version, fragment hits were clustered to incorporate the possibility of multiple distinct fragments being found within a single protein domain. If standard cutoffs are applied, we still observe the same power-law distribution of domain connectivity, with few domains accumulating most of the network's links in a highly populated major component (Ferruz et al., 2021).

A major improvement is the addition of ligand information (**Supplementary Figure S1**). It is now possible to search for ligands in two ways: either by its PDB code (e.g., adenosine-5′-triphosphate: ATP) or by its SMILES. SMILES searches not only provide identical or 70% similar ligands, but also superstructures or substructures of the compound using Tanimoto coefficients. To cite an example, searching for substructures of ATP would also provide all fragments that bind substructures of it, like adenosine or inorganic phosphate.

In addition, we also enable visualization of networks of proteins bound to certain ligands and provide this information in a downloadable table. The table includes not only the statistics of the fragment and the most connected entry as a representative but also all ligands that are bound to the fragment. It is possible to retrieve additional information for each ligand and to directly visualize the protein-ligand interactions in the context of the selected fragment using the NGL Viewer. One can toggle interactions like π–π stacking, hydrogen bonds, compute distances, and show surface representations.

A PBP-like Conserved Fragment

Fuzzle 2.0's new features enhance the analysis of fragments. Here we want to illustrate them by exploring the evolutionary relationship of a member of the periplasmic binding protein-like I fold (PBP-like I, c.93) (**Figure 1A**). In Fuzzle we also find hints of these evolutionary relationships between PBPs and other folds. Here, we use the ribose binding protein from *Thermotoga maritima* (tmRBP, d2fn9a_). It belongs to the PBP-like I superfamily c.93.1 (Cuneo et al., 2008). We observe that d2fn9a_ appears in an unusually high number of hits either as query or subject (altogether 2028) with other protein domains in the database without any cutoffs, as can be queried with the software Protlego (Ferruz et al., 2021). 1,566 of those hits correspond to local alignments shared with domains that belong to other superfamilies and folds. This domain is more connected than observed for the average domain in Fuzzle (172 hits/domain). If we focus on standard cutoffs, we obtain 121 hits, belonging to 15 superfamilies and 9 folds. These numbers indicate that domain d2fn9a_ shares several conserved fragments with other domains.

This high evolutionary connectivity can be viewed in Fuzzle when looking at the HHsearch hits to sequences of other superfamilies (**Figure 1B**). tmRBP shows an unusual number of local alignments in its N-terminal region, indicating a

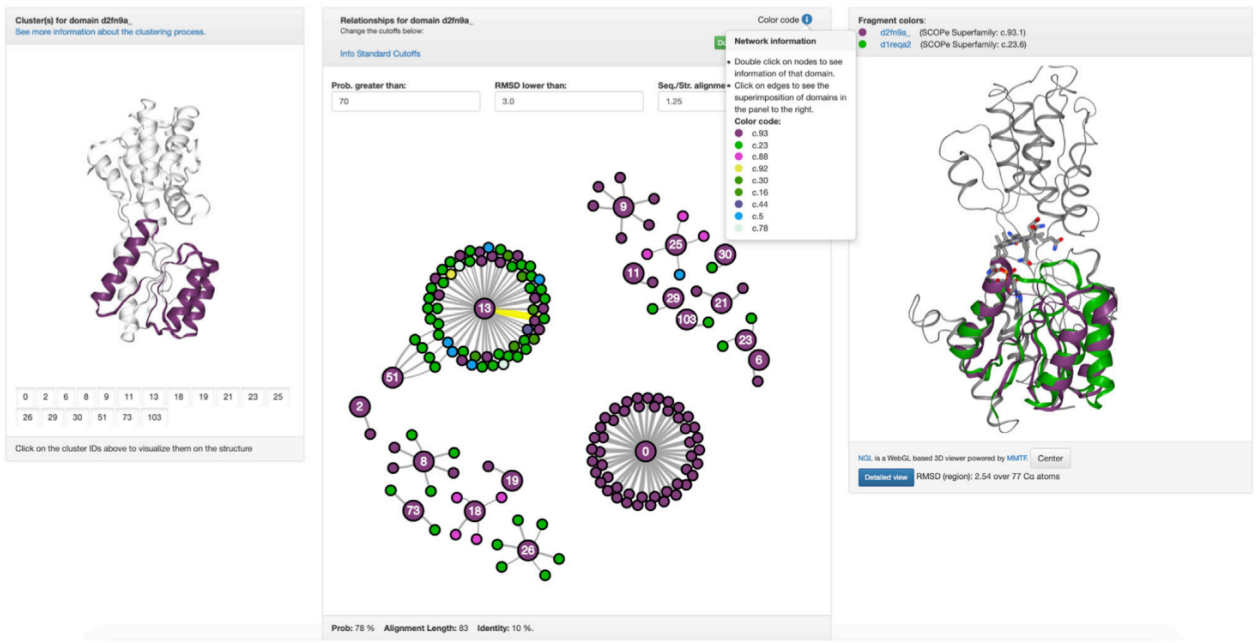


FIGURE 2 | Conservation in tmRBP. (c.f. https://fuzzle.uni-bayreuth.de/2.0/hh/fragment_graph/d2fn9a_/). **Left:** The panel highlights each hit in d2fn9a_. In this case, cluster 13 has been selected, the main object of this study. **Middle:** The clusters of d2fn9a_ are shown, where each interactive node is a domain, linked to other domains when they share a fragment. Each 'island-like' cluster corresponds to a set of proteins that have a fragment in common. Nodes are colored according to their folds **(top)**. Mousing over an edge it gets highlighted (yellow) and the alignment parameters in the footer are shown **(bottom)**. An individual PyMOL session of superposed structures for each cluster can be downloaded **(top green button)**. **Right:** Upon clicking on an edge between two nodes **(middle panel)** the superposition of the structures with their fragment colored according to their fold will be shown on the right.

conserved fragment. We thus decided to look in detail at this fragment and characterize it. This is possible by using domain-centered networks in Fuzzle (**Figure 2**). In this representation, the domain in question is defined as an interactive circle, and other domains that have fragments in common are linked to it. d2fn9a_ always appears as the center of each 'island' or cluster. In this representation, we show all hits that surpass the previously described cutoffs but unlike **Figure 1B** it includes hits with domains from the same superfamily as well. The ID numbers (**Figure 2**, left) are not contiguous since not all fragments fulfill the user-defined cutoffs. To discern them from the previously defined fragments, we will call these connections within the domain-centered networks 'clusters'. Clicking on the identifiers to the left depicts the cluster in the domain (**Figure 2**, left).

For tmRBP, Fuzzle identifies a total of 153 hits to 136 other domains (**Figure 2**, **Supplementary Table S1**) using the standard cutoffs. These hits can then be grouped into 18 clusters that map to different regions of tmRBP (**Figure 2**, **Supplementary Table S2**). The coloring scheme matches throughout Fuzzle 2.0 representing the individual folds. For example, sequences that are shown in green in **Figure 1B** also appear as green nodes in **Figure 2**. We can thus infer that these sequences have cluster 13 in common. The fragment position confirms this observation (positions 11–87, **Supplementary Table S2**, **Figure 2**). With 63 domains in this subgraph, cluster 13 constitutes d2fn9a_'s most promiscuous fragment. Structurally, it contains the three N-terminal helices and four β-sheets, with the first β-sheet not necessarily present in all domains. In total, cluster 13 spans domains from 8 folds to 12 superfamilies: c.16.1 (Lumazine synthase): 4 domains, c.23.1 (CheY-like): 2 domains, c.23.13 (Type II 3-dehydroquinate dehydratases): 9 domains, c.23.16 (Class I glutamine amidotransferase-like): 9 domains, c.23.6 (Cobalamin (vitamin B_{12})-binding domain), c.23.8 (N^5-CAIR mutase (Phosphoribosylaminoimidazole carboxylase, PurE): 2 domains, c.30.1 (PreATP-grasp domain): 11 domains, c.44.3 (PIWI domain N-terminal-like): 1 domain, c.5.1 (c.5.1: MurCD N-terminal domain): 3 domains, c.78.2 (Aspartate/glutamate racemase): 2 domains, c.92.3 (PrpR receptor domain-like): 1 domain, and c.93.1 (Periplasmic binding protein-like I): 16 domains.

Ligand-Binding in the Conserved PBP-like Fragment

One of the major goals of Fuzzle 2.0 is not only to update our platform for evolutionary analyses but also to facilitate searches for suitable fragments to design protein chimeras. In the case of d2fn9a_, parts of the protein that correspond to cluster 13 could be replaced with a homologous and structurally well-superimposed fragment of another protein that possesses an interesting function (Ferruz et al., 2021). d2fn9a_'s cluster 13 is a good candidate for this study, as it contains more than 63 direct hits in the same and other superfamilies that contain the same structural fragment with large deviations in sequence and

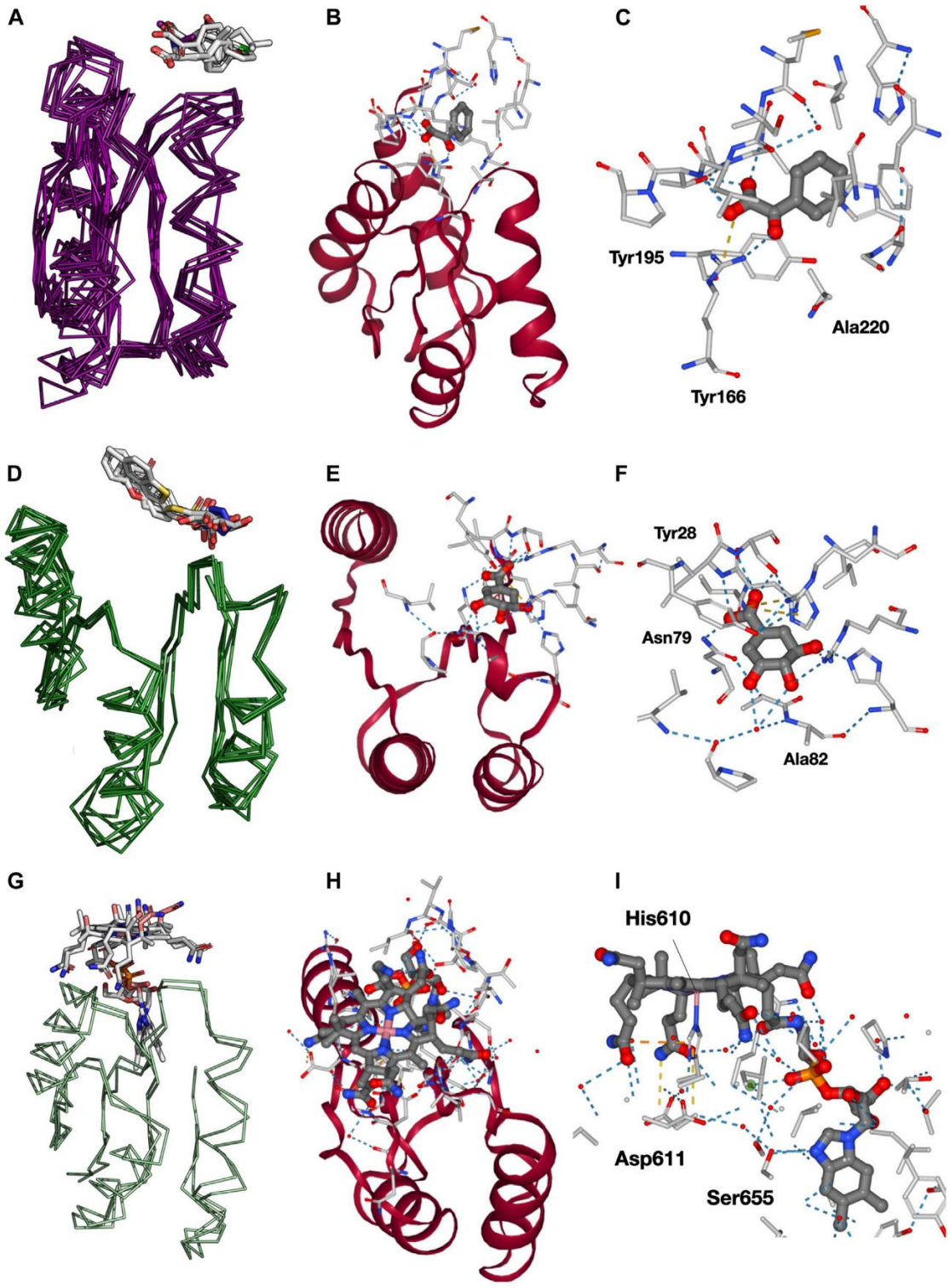


FIGURE 3 | Analysis of binding modes in cluster 13 for superfamily c.93.1, c.23.13, and c.23.6. **(A)** Domains in c.93.1 that contain cluster 13 and bind a ligand. **(B)** Example of tyrosine (Tyr) bound to domain d3snra_ and its interactions with cluster 13 (red). **(C)** Closer view onto the binding pocket (https://fuzzle.uni-bayreuth.de/2.0/ligands/detailed/d3td9a_/135/219/. **(D)** Domains in c.23.13 that contain cluster 13 and bind a ligand. **(E)** 3-dehydroshikimate (DHK) bound to domain d1gtza_ and its interactions with cluster 13 (red). **(F)** Closer view on the binding pocket (https://fuzzle.uni-bayreuth.de/2.0/ligands/detailed/d1gtza_/6/81). **(G)** The two domains in c.23.6 that contain cluster 13 and bind a ligand. **(H)** Cobalamin (B12) bound to domain d1reqa2_ and its interactions with cluster 13 (red). **(I)** Closer view onto the binding pocket. Most of B12 interactions are performed within the fragment (https://fuzzle.uni-bayreuth.de/2.0/ligands/detailed/d1reqa2/598/687).

TABLE 1 | List of domains containing cluster 13 with bound ligands.

Domain	PDB code	Superfamily	Compound name
d2obxa_	INI	c.16.1	5-nitro-6-ribityl-amino-2,4 (1 h,3 h)-pyrimidinedione
d1gtza_	DHK	c.23.13	3-dehydroshikimate
d2y71a_	CB6	c.23.13	(1r,4s,5r)-1,4,5-trihydroxy-3-[(5-methyl-1-benzothiophen-2-yl)methoxy]cyclohex-2-ene-1-carboxylic acid
d2c4wa_	GAJ	c.23.13	N-tetrazol-5-yl 9-oxo-9h-xanthene-2 sulphonamide
d5ydba_	DQA	c.23.13	3-dehydroquinic acid
d2xdaa_	JPS	c.23.13	(4r,6r,7s)-2-(2-cyclopropylethyl)-4,6,7-trihydroxy-4,5,6,7-tetrahydro-1-benzothiophene-4-carboxylic acid
d1a9xb2	CYG	c.23.16	2-amino-4-(amino-3-oxo-propylsulfanylcarbonyl)-butyric acid
d1reqa2	B12	c.23.6	Cobalamin
d1ccwa_	CNC	c.23.6	Co-cyanocobalamin
d2atea_	NIA	c.23.8	4-nitro-5-aminoimidazole ribonucleotide
d2bgga2	U	c.44.3	Uridine-5'-monophosphate
d2x5oa1	VSV	c.5.1	N-({3-[({4-[(z)-(2,4-dioxo-1,3-thiazolidin-5-ylidene)methyl]phenyl}amino)methyl]phenyl}carbonyl)-D-glutamic acid
d1p3da1	UMA	c.5.1	Uridine-5'-diphosphate-n-acetylmuramoyl-l-alanine
d3t23a_	TYR	c.93.1	Tyrosine
d3td9a_	PHE	c.93.1	Phenylalanine
d4nqra_	ALA	c.93.1	Alanine
d3sg0a_	173	c.93.1	Benzoyl-formic acid
d4q6ba_	LEU	c.93.1	Leucine
d3ipca1	LEU	c.93.1	Leucine
d4n0qa_	LEU	c.93.1	Leucine
d3snra_	TYR	c.93.1	Tyrosine

function. These deviations, although large, are still remnants of a remote homologous ancestor, but more importantly, provide a wide range of functionalities that we can exploit for protein design purposes. In this example, we have looked at the ligand-binding capabilities of the 63 domains containing d2fn9a_'s cluster 13. These fragments, provided there are no structural clashes, could hence be potential candidates for replacement as previously achieved experimentally with the PBP-flavodoxin-like chimera (PDB id: 4QWV). Besides, they could represent a starting point for protein engineering, especially those fragments that entirely encapsulate a ligand offering an opportunity for binding site transfer. This can for example be done with the recently published Protlego tool (Ferruz et al., 2021). Here, we have characterized the ligand-binding proteins containing cluster 13 and analyzed their prospects for protein engineering.

To this end, we downloaded the PyMOL session with a superposition of all cluster 13-containing domains, available from the domain-centered network view for all fragments (**Figure 3**, middle). Despite their large sequence divergence, the backbone of the structures is quite conserved (**Supplementary Figure S2**). In the superposition, we observe several ligands bound to the fragment, mostly at the top position, and some other ligands and solvents interacting with different regions of the protein. We noticed that several of these ligands correspond to additives commonly found in crystallization media, which were discarded from our analysis (see Methods). Ligands within 4 Å of any heavy atom of cluster 13 are summarized in **Table 1**. Not all superfamilies containing cluster 13 are shown, since from the 12 superfamilies only 8 have ligands bound, and only 4 have 2 or more representatives.

Naturally, the most abundant superfamily is the one of the ribose binding proteins itself (c.93.1, **Figure 3A**), with 8 domains (**Table 1**, **Supplementary Figure S3**). Visualization of these domains reveals that they mostly bind amino acid ligands in a conserved binding mode (**Figure 3B**), along with benzoyl-formic acid (PDB ligand code 173). As in most PBPs, all ligands in this set bind in the cleft defined by the two protein lobes. Because cluster 13 is located in the N-terminal lobe, it only interacts *via* a few residues with the respective ligand. Two of these interactions are particularly conserved among the sequences, a Tyr or Phe residue, and a Ser, Tyr, or Ala residue (**Supplementary Figure S3**). We particularly looked at the interactions with Fuzzle 2.0's detailed viewer for domain d3sg0a_ (https://fuzzle.uni-bayreuth.de/2.0/ligands/detailed/d3sg0a_/139/236/), which contains overall 3 interactions with the fragment: the conserved residues Tyr168 and Ala220 (**Supplementary Figure S3**) that interact *via* hydrophobic packing and Arg195 that forms a salt bridge with the ligand (**Figure 3C**). Particularly important is the salt bridge formed between Arg195's guanidium group and ligand 173's carboxyl, an interaction that also appears in two other domains, namely d3snra_ and d3t23a_ (note, that the corresponding PDB entries have been superseded by 3UK0 and 3UK1, respectively).

The second most abundant superfamily is the Type II 3-dehydroquinate dehydratase superfamily (c.23.13), including 5 ligand-binding domains (**Supplementary Figure S3**). These domains bind ligands that are artificial drugs used as antimicrobial agents (**Figure 3D**). **Figure 3E** shows the ligand DHK (3-dehydroshikimate) binding domain d1gtza_ interacting with additional residues outside the fragment. However, two critical interactions are contained in the fragment (Ala82 and Asn79). These interactions are highly conserved among all c.23.13 sequences (**Supplementary Figure S3**).

The cobalamin (vitamin B_{12})-binding superfamily (c.23.6) is represented by two domains in the receptor set (**Supplementary Figure S3**). Both domains bind cobalamin variants (cobalamin and co-cyanocobalamin) in a conserved fashion (**Figure 3G**).

Figure 3H shows domain d1reqa2 in complex with B_{12} (cobalamin). The ligand performs most of its interactions with the fragment (**Figure 3I**), especially with the loop between β1 and α1 with the cobalt-coordinating His610. Other important residues are Asp611 (loop1), and Ser655 (β2; https://fuzzle.uni-bayreuth.de/2.0/ligands/detailed/d1reqa2/598/687).

Other less abundant superfamilies binding ligands are c.5.1, c.16.1, c.44.3, and c.23.16, shown in **Supplementary Figure S4**. Interestingly, superfamily c.5.1 uses a different mode of binding, with the ligand bound between helices. A superposition with domain d2fn9a_ reveals a different topology where the β-sheets do not exactly superimpose, leaving β3 for d2fn9a_ unmatched. Superfamily c.23.16 contains only one domain with chemical entities, which however only interact with residues mostly outside cluster 13 of domain d1a9xb2 (**Supplementary Figure S4B**). Superfamily c.44.3 with its representative d2bgga2 binds two molecules of uridine-5'-monophosphate (**Supplementary Figure S4C**). The ligand binds at the edge of d2bgga2's β4, which also corresponds to the terminal part of the domain. Superfamily c.16.1 is represented by domain d2obxa_, which binds ligand 5-nitro-6-ribityl-amino-2,4 (1 h, 3 h)-pyrimidinedione (NRP, PDB ligand code INI). All interactions for this ligand are contained in the fragment. These examples show that protein ligand interactions often occur at similar positions in a protein corresponding to the fragments detected in Fuzzle. However, the mode of binding of various ligands in different homologous proteins may vary, e.g. as described for the superfamily c.5.1.

DISCUSSION

We recently published the Fuzzle (Fold Puzzle) database which contains a set of evolutionarily related protein fragments that can also be used for protein design. It is known that modern proteins evolved by replicating and recombining smaller sequence fragments. Fuzzle offers the opportunity to identify these fragments and to mimic evolutionary processes in the lab as well as to build new proteins. In this updated version of Fuzzle 2.0, we enhanced the analysis tools and extended them to include detailed information about protein fragment-ligand interactions. This extension now enables the identification and analysis of ligands and their interactions with a conserved fragment. As a note of caution: since Fuzzle is based on a non-redundant dataset of SCOPe, some ligand information might be incomplete. Thus, it will be still necessary to consult the literature or other databases for an in-depth analysis of a specific protein-ligand interaction.

Using a periplasmic binding protein (PBP) fragment as an example, we demonstrated the new features of Fuzzle 2.0. SCOP places this fold into a single superfamily, periplasmic binding protein-like I (PBP-like I superfamily, c.93.1). Its fold consists of two similar intertwined lobes with 3 layers (α/β/α), each composed of a parallel six-stranded β-sheet with the order 213456 (**Figure 1A**). There also exists a two-lobed PBP-like II fold (c.94), with a somewhat different topology. The two lobes in both folds define a small hinge region that recognizes a large number of ligands and ions in bacteria. PBPs exist in an open and closed conformation, with the open conformation predominating in the absence of ligands (Kröger et al., 2021). Such conformational plasticity have led to PBPs being widely used in biosensing applications (Grünewald, 2014). Based on structural consideration alone it has long been proposed that the PBP-like I fold arose *via* gene duplication from a flavodoxin-like fold (c.23) (Louie, 1993; Fukami-Kobayashi et al., 1999). In fact, a protein chimera could be built through combination of fragments from these two folds (PDB id: 4QWV). A similar postulated duplication event has also recently been explored for the emergence of the two-lobed HemD-like fold from flavodoxin-like proteins (Toledo-Patiño et al., 2019), combining sequence and structural analysis with experimental reconstruction.

Here, we have focused our analysis on the ribose binding protein from *Thermotoga maritima* (tmRBP), a single domain protein (d2fn9a_). The domain contains many sequence similarities to other superfamilies, especially in its N-terminal region (**Figure 1B**). This region corresponds to a conserved fragment spanning 3 helices and four β-strands (**Figure 2**). The fragment occurs in 63 domains of 12 different superfamilies, and thus offers a great prospect for protein engineering. A detailed protein-ligand analysis was described for 22 of the domains, distributed over seven of the 12 identified superfamilies (**Table 1**).

The dataset of ligand-binding domains offers opportunities for protein design. While the engineering of ligand-binding pockets has become more successful over the years, it is still difficult. Now, reusing ready-made parts from existing proteins can help overcome some of the difficulties. Therefore, we suggest chimeragenesis by replacement in which the corresponding fragment in d2fn9a_ gets replaced by a homologous fragment binding a ligand such as the one in INI-binding d2obxa_ domain. Such an approach has been successfully applied in several instances and offers a novel route for functional diversification (Lechner et al., 2018). Another interesting opportunity that this approach offers is to test evolution by protein engineering as was previously shown for the HemD fold, another bilobular protein. The protein could be dissected into its two lobes, one of which was shown to fold by itself into the related flavodoxin-like fold c.23 (Toledo-Patiño et al., 2019). We would expect similar behaviour for the PBP-like folds.

One question that remains is whether the lower PBP-lobe could adopt the functionality of some of its related proteins like those described above belonging to the flavodoxin-like fold. Here, we have shown that domains of several superfamilies (c.93.1, c.23.6, and c.5.1) bind different ligands at similar regions in the protein structure; however, the mode of binding can differ. This region represents a conserved fragment and therefore strengthens the hypothesis that domains contain conserved building blocks even shared by seemingly unrelated folds. This observation gives rise to the possibility to identify potential ligands that could bind to a domain. In the example described the analysis suggests that the identified fragment in d2fn9a_ could be capable of recognizing ligands B12 or INI after performing several rounds of mutations either by protein engineering or directed evolution.

Overall, we believe that the new version of Fuzzle will be a valuable tool for various fields of research. On the one hand Fuzzle 2.0 allows evolutionary biologists to strengthen the evidence for common ancestry and on the other hand allows protein designers to use this information in transferring ligand binding sites into other protein scaffolds.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://fuzzle.uni-bayreuth.de/2.0.

AUTHOR CONTRIBUTIONS

NF, FM, FL, SS, and BH contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

FUNDING

We acknowledge financial support from the European Research Council (ERC Consolidator grant 647548 "Protein Lego") and VolkswagenStiftung (grant 94747).

ACKNOWLEDGMENTS

We thank Johannes König for his help in expanding the web server.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.715972/full#supplementary-material

REFERENCES

Alva, V., Söding, J., and Lupas, A. N. (2015). A Vocabulary of Ancient Peptides at the Origin of Folded Proteins. *Elife* 4, e09410. doi:10.7554/eLife.09410.001

Alva, V., and Lupas, A. N. (2018). From Ancestral Peptides to Designed Proteins. *Curr. Opin. Struct. Biol.* 48, 103–109. doi:10.1016/j.sbi.2017.11.006

Bharat, T. A. M., Eisenbeis, S., Zeth, K., and Höcker, B. (2008). A -barrel Built by the Combination of Fragments from Different Folds. *Proc. Natl. Acad. Sci.* 105, 9942–9947. doi:10.1073/pnas.0802202105

Cuneo, M. J., Beese, L. S., and Hellinga, H. W. (2008). Ligand-induced Conformational Changes in a Thermophilic Ribose-Binding Protein. *BMC Struct. Biol.* 8, 50. doi:10.1186/1472-6807-8-50

Farías-Rico, J. A., Schmidt, S., and Höcker, B. (2014). Evolutionary Relationship of Two Ancient Protein Superfolds. *Nat. Chem. Biol.* 10, 710–715. doi:10.1038/nchembio.1579

Ferruz, N., Lobos, F., Lemm, D., Toledo-Patino, S., Farías-Rico, J. A., Schmidt, S., et al. (2020). Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design. *J. Mol. Biol.* 432, 3898–3914. doi:10.1016/j.jmb.2020.04.013

Ferruz, N., Noske, J., and Höcker, B. (2021). Protlego: a Python Package for the Analysis and Design of Chimeric Proteins. *Bioinformatics*, btab253. doi:10.1093/BIOINFORMATICS/BTAB253

Fox, K. N., Brenner, E. S., and Chandonia, J. M. (2014). SCOPe: Structural Classification of Proteins—Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures. *Nucleic Acids Res.* 42, D304–D309. doi:10.1093/nar/gkt1240

Fukami-Kobayashi, K., Tateno, Y., and Nishikawa, K. (1999). Domain Dislocation: A Change of Core Structure in Periplasmic Binding Proteins in Their Evolutionary History. *J. Mol. Biol.* 286, 279–290. doi:10.1006/jmbi.1998.2454

Godden, J. W., Xue, L., and Bajorath, J. (2000). Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* 40, 163–166. doi:10.1021/ci990316u

Goncearenco, A., and Berezovsky, I. N. (2015). Protein Function From Its Emergence to Diversity in Contemporary Proteins. *Phys. Biol.* 12, 045002. doi:10.1088/1478-3975/12/4/045002

Grünewald, F. S. (2013). Periplasmic Binding Proteins in Biosensing Applications. *Bioanal. Rev.* 1, 205–235. doi:10.1007/11663_2013_7

Horowitz, N. H. (1945). On the Evolution of Biochemical Syntheses. *Proc. Natl. Acad. Sci.* 31, 153–157. doi:10.1073/pnas.31.6.153

Jensen, R. A. (1976). Enzyme Recruitment in Evolution of New Function. *Annu. Rev. Microbiol.* 30, 409–425. doi:10.1146/annurev.mi.30.100176.002205

Kröger, P., Shanmugaratnam, S., Ferruz, N., Schweimer, K., and Höcker, B. (2021). A Comprehensive Binding Study Illustrates Ligand Recognition in the Periplasmic Binding Protein PotF. *Structure* 29, 433–443. doi:10.1016/j.str.2020.12.005

Laurino, P., Tóth-Petróczy, Á., Meana-Pañeda, R., Lin, W., Truhlar, D. G., and Tawfik, D. S. (2016). An Ancient Fingerprint Indicates the Common Ancestry of Rossmann-fold Enzymes Utilizing Different Ribose-Based Cofactors. *PLOS Biol.* 14, e1002396. doi:10.1371/journal.pbio.1002396

Lechner, H., Ferruz, N., and Höcker, B. (2018). Strategies for Designing Non-natural Enzymes and Binders. *Curr. Opin. Chem. Biol.* 47, 67–76. doi:10.1016/j.cbpa.2018.07.022

Longo, L. M., Jabłońska, J., Vyas, P., Kanade, M., Kolodny, R., Ben-Tal, N., et al. (2020). On the Emergence of P-Loop Ntpase and Rossmann Enzymes from a Beta-Alpha-Beta Ancestral Fragment. *Elife* 9, 1–16. doi:10.7554/ELIFE.64415

Louie, G. V. (1993). Porphobilinogen Deaminase and its Structural Similarity to the Bidomain Binding Proteins. *Curr. Opin. Struct. Biol.* 3, 401–408. doi:10.1016/S0959-440X(05)80113-7

Narunsky, A., Kessel, A., Solan, R., Alva, V., Kolodny, R., and Ben-Tal, N. (2020). On the Evolution of Protein-Adenine Binding. *Proc. Natl. Acad. Sci. USA* 117, 4701–4709. doi:10.1073/pnas.1911349117

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2017). Complex Evolutionary Footprints Revealed in an Analysis of Reused Protein Segments of Diverse Lengths. *Proc. Natl. Acad. Sci. USA* 114, 11703–11708. doi:10.1073/pnas.1707642114

Ohta, T. (2000). Mechanisms of Molecular Evolution. *Phil. Trans. R. Soc. Lond. B* 355, 1623–1626. doi:10.1098/rstb.2000.0724

Romero Romero, M. L., Yang, F., Lin, Y.-R., Toth-Petroczy, A., Berezovsky, I. N., Goncearenco, A., et al. (2018). Simple yet Functional Phosphate-Loop Proteins. *Proc. Natl. Acad. Sci. USA* 115, E11943–E11950. doi:10.1073/pnas.1812400115

Romero-Romero, S., Kordes, S., Michel, F., and Höcker, B. (2021). Evolution, Folding, and Design of TIM Barrels and Related Proteins. *Curr. Opin. Struct. Biol.* 68, 94–104. doi:10.1016/j.sbi.2020.12.007

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlić, A., and Rose, P. W. (2018). NGL Viewer: Web-Based Molecular Graphics for Large Complexes. *Bioinformatics* 34, 3755–3758. doi:10.1093/bioinformatics/bty419

Schmidt, S., and Dandekar, T. (2002). *Gene Regulations and Metabolism - Postgenomic Computational Approaches*. Editor J. Collado-Vides and R. Hofestädt (MIT Press Cambridge, Massachusetts London, England).

Sikosek, T., and Chan, H. S. (2014). Biophysics of Protein Evolution and Evolutionary Protein Biophysics. *J. R. Soc. Interf.* 11, 20140419. doi:10.1098/rsif.2014.0419

Söding, J. (2005). Protein Homology Detection by HMM-HMM Comparison. *Bioinformatics* 21(7), 951–960. doi:10.1093/bioinformatics/bti125

Toledo-Patiño, S., Chaubey, M., Coles, M., and Höcker, B. (2019). Reconstructing the Remote Origins of a Fold Singleton from a Flavodoxin-like Ancestor. *Biochemistry* 58, 4790–4793. doi:10.1021/acs.biochem.9b00900

Weininger, D. (1988). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* 28, 31–36. doi:10.1021/ci00057a005

Yang, J., Roy, A., and Zhang, Y. (2013). BioLiP: A Semi-manually Curated Database for Biologically Relevant Ligand-Protein Interactions. *Nucleic Acids Res.* 41, D1096–D1103. doi:10.1093/nar/gks966

Zhang, Y., and Skolnick, J. (2005). TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* 33(7), 2302–2309. doi:10.1093/nar/gki524

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supporting Information for:

Fuzzle 2.0: Ligand Binding in Natural Protein Building Blocks

Noelia Ferruz[1*], Florian Michel[1], Francisco Lobos[1], Steffen Schmidt[2], Birte Höcker[1*]
[1]Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany,
[2]Computational Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany.
*** Correspondence:**
Corresponding Author
birte.hoecker@uni-bayreuth.de, noelia.ferruz-capapey@uni-bayreuth.de
Keywords: web server, protein evolution, protein design, protein fragment, flavodoxin-like fold, periplasmic binding protein.

Table S1: Statistics of Datasets. Fuzzle 2.0 allows to access two datasets, SCOP 2.06 and SCOP 2.07. The last row reports the pairwise hits using the filtering criteria reported in the manuscript (HHSearch Probability > 70, TM-score > 0.3 with at least 10 C_α-Atoms superposed, a RMSD below 3.0 Å, and a ratio between the sequence and structure lengths of maximum 1.25.

Dataset	SCOP 2.06	SCOP 2.07
Families	4,783	4,849
Superfamilies	2,006	2,024
Folds	1,221	1,232
Fuzzle hits	8,109,195	10,434,359
Fuzzle hits (filtered)	4,970,087	6,255,666

Table S2: The 18 c**lusters found in the ribose binding protein**. The cluster identifiers correspond to Figure 3. The start/end positions match the amino acid sequence of d2fn9a_. The number of domains that are contained in each cluster is shown in the last column. Note, that multiple fragments can be found within a single domain but are sorted into different clusters, e.g. 4 domains are found in both cluster 51 and 13 and therefore are counted twice, resulting in a greater number of total reported domains.

Cluster	Start	End	Domains
0	2	280	49
2	2	223	2
6	5	135	2
8	2	123	7
9	129	264	7
11	112	252	2
13	11	87	63
18	47	115	5
19	2	116	2
21	2	101	3
23	169	256	3
25	44	87	5
26	17	106	7
29	149	234	2
30	53	105	2
51	15	93	5
73	152	257	3
103	22	120	2

Figure S1: Most common ligands found in conserved fragments.

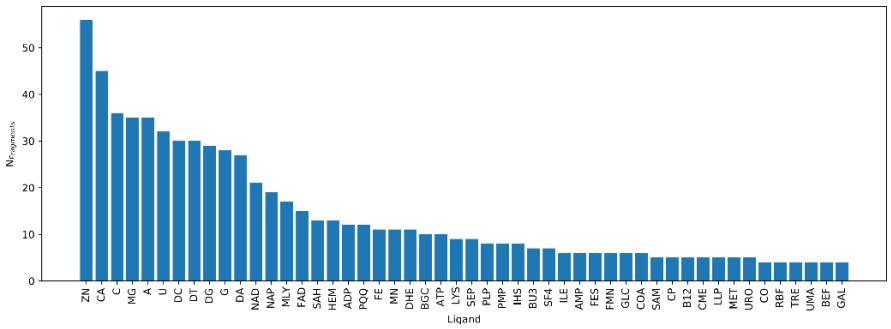


Figure S2: Superposition of all domains that contain cluster 13 (https://fuzzle.uni-bayreuth.de/2.0/super/pymol/cluster/d2fn9a__13/70/3.0/1.25).

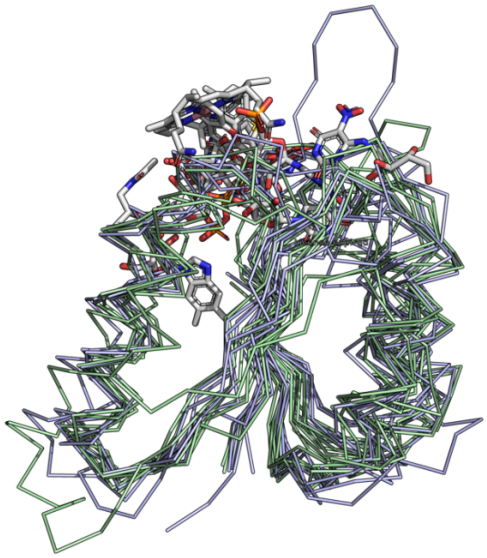


Figure S3: Sequence alignment for the ligand-binding domains in superfamilies c.93.1, c.23.13, and c.23.6. Interactions with the ligands are highlighted.

(a) c.93.1: Periplasmic binding protein-like I

```
d3snra_   -GYIGYSDSYGDLWFNDLKKQGEAMGLKIVGEERFARPDTSVAGQALKLVAANPDAILVGASGTAAALPQTTLRE-RGYNGLIYQTHG
d3t23a_   -GYIGYSDSYGDLWFNDLKKQGEAMGLKIVAEERFARPDTSVAGQVLKLVAANPDAILVGASGTAAALPQTALRE-RGYNGLIYQ---
d3sg0a_   -GYIGFSDAYGEGYYKVLAAAAPKLGFELTTHEVYARSDASVTGQVLKIIATKPDAVFIASAGTPAVLPQKALRE-RGFKGAIYQ---
d3ipca1   -AIIHDKTPYGQGLADETKKAANAAGVTEVMYEGVNVGDKDFSALISKMKEAGVSIIYWGGLHTEAGLIIRQAAD-QGLKAKLVS---
d4n0qa_   -AVIHDKGAYGKGLADAFKAAINKGGITEVHYDSVTPGDKDFSALVTKLKSAGAEVVYFGGYHAEGGLLSRQLHD-AGMQALVLG---
d3td9a_   -VFTDVEQDYSVGLSNFFINKFTELGG-QVKRVFFRSGDQDFSAQLSVAMSFNPDAIYITGYYPEIALISRQARQ-LGFTGYILA---
d4q6ba_   -VIYYTDDSYGNGLANAFEDYARAQGITIVDRFNYYGNLKDLERLYDKWQAFGMDGIFIAKTATGGGTEFLVDAKSVGIEVPLIA---
d4nqra_   AVFFAQNDAFSKSETEIFQQTVKDQGLELVTVQKFQTTDTDFQSQATNAINLKPDLVIISGYAADGGNLVRQLRE-LGYQGAIIG---
```

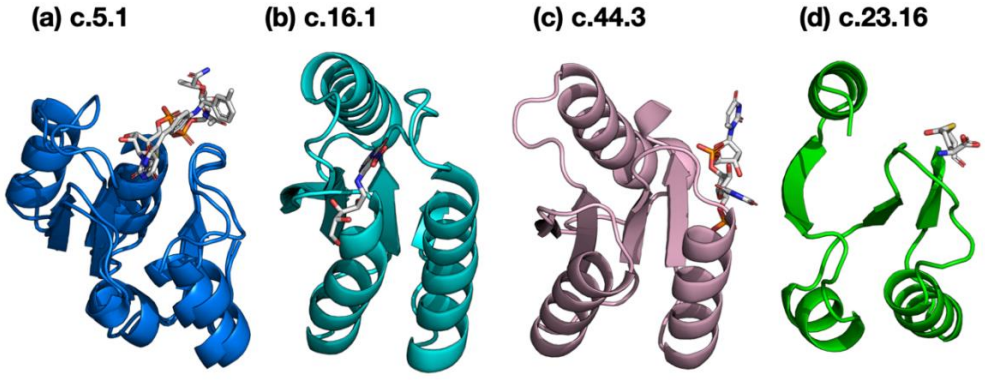
(b) c.23.13: Type II 3-dehydroquinate dehydratase

```
d2c4wa_    QIHEIMQTFVKQGNLDVELEFFQTNFEGEIIDKIQESVGSEYEGIIINPGAFSHTSIAIADAIMLAG-KPVIEVH
d2xdaa_    QIHEIMQTFVKQGNLDVELEFFQTNFEGEIIDKIQESVGSDYEGIIINPGAFSHTSIAIADAIMLAG-KPVIEVH
d5ydba_    --NINRQLIAQAEQASITLDTFQSNWEGAIVDRIHQAQTEGVKLIIINPAALTHTSVALRDALLGVA-IPFIEVH
d1gtza_    --DVEALCVKAAAAHGGTVDFRQSNHEGELVDWIHEAR-LNHCGIVINPAAYSHTSVAILDALNTCDGLPVVEVH
d2y71a_    --ELVALIEREAAELGLKAVVRQSDSEAQLLDWIHQAA-DAAEPVILNAGGLTHTSVALRDACAELS-APLIEVH
```

(c) c.23.6: Cobalamin (vitamin B12)-binding domain

```
d1reqa2    RILLAKMGQDGHDRGQKVIATAYADLGFDVDVGPLFQTPEETARQAVEADVHVVGVSSLAGGHLTLVPALRKELDKLGRPDILITVGGV
d1ccwa_    TIVLGVIGSDCHAVGNKILDHAFTNAGFNVVNIGVLSPQELFIKAAIETKADAILVSSLYGQGEIDCKGLRQKCDEAGLEGILLYVGGN
```

Figure S4: Ligand-binding domains containing cluster 13 from less numerous superfamilies.



7. Paper IV

Michel, F., Kossendey T., Höcker, B.
Isolation of subdomain-sized elements in a modern periplasmic binding protein.
Manuscript

Isolation of subdomain-sized elements in a modern periplasmic binding protein

Florian Michel,[1†] Timo Kossendey,[1†] Birte Höcker[1*]

[1] Department of Biochemistry, University of Bayreuth, Bayreuth 95447, Germany.

[†] These authors contributed equally to the work.

Correspondence

* Corresponding author. Birte Höcker. Department of Biochemistry, University of Bayreuth, Bayreuth 95447, Germany. Phone: +490921557845. **E-mail:** birte.hoecker@uni-bayreuth.de

ORCID identifiers

Florian Michel: 0000-0002-5111-8290

Birte Höcker: 0000-0002-8250-9462

Keywords

Protein evolution, periplasmic binding protein, solute binding protein, ribose binding protein, protein evolution, protein fragment

This PDF file includes:

- Main Text.
- Figures 1-3.
- Table 1.

Abstract

One of the core questions in investigating the evolution of proteins is the genesis of the protein structural universe that we see today. It is generally believed that the modern diversity of protein arose from the coalescence of an ancestral set of small subsets of polypeptide fragments. Through the implementation of increasingly sensitive bioinformatic methods several datasets emerged recently to classify these remnants of this sub-domain regime. Using the web-based tool Fuzzle (Fold Puzzle Database; *https://fuzzle.uni-bayreuth.de*) (Ferruz, 2020), we identified a candidate for such a remnant fragment in a modern periplasmic-binding protein. The analyzed consensus fragments as well as the sequence taken directly from the parental protein were then overexpressed. We found the fragments to fold in solution, and mostly adopt a dimeric conformation. These findings significate that while not necessarily carrying out any essential function, these fragments are not just folded and stable in isolation, but also significantly resistant to changes in their sequence.

Main Text

Introduction

Most molecular mechanisms in modern cells are carried out by proteins. This complex machinery allowed life to adapt to different environments. However, in contrast to the many functions proteins carry out, their structural complexity is relatively limited. The domain has long been regarded as the commonly shared, independently folding unit within folds, which have been reused and adapted by nature. Many de-facto standards of structural protein classification is done via sequence homology of domains, and catalogued in well-known databases like SCOP, ECOD or CATH (Andreeva, 2014; Cheng, 2014 ;Sillitoe, 2018). However, at least from an evolutionary point of view they are no longer to be considered the smallest defined building block in proteins. Recent research on the structure and sequence of

proteins using modern bioinformatic methods has shown that there is a shared set of sub-domain fragments not just within, but between protein folds (Höcker, 2014; Alva, 2015; MacKenzie, 2016; Nepomnyachiy, 2017; Ferruz, 2020; Konagurthu, 2021). A possible explanation for this is the idea that the modern protein universe started off with a limited subset of smaller, independent fragments (Alva, 2009). Multiplication, rearrangement, and fusion of these fragments then led to the creation of bigger proteins, which could enact the more complex functions needed for more elaborate life to exist on earth (Ohta, 2000). The fact that several of these fragments can still be observed to be shared between folds would imply their existence before the divergence of these folds. In previous research we used the *Fuzzle* database to identify such a fragment in the ribose-binding protein of *Thermotoga maritima* (RBP) and explored its evolutionary relationship between its PBP-like fold and other folds (Ferruz, 2021).

To understand what makes this conserved N-terminal fragment so significant to be shared between so many folds, we analyzed the corresponding sequence of *T. maritima* RBP (residues 1-88) in isolation of its structural context. Additionally, to generate a preliminary idea on how changes in the sequence of this fragment could influence its behavior, different consensus sequences of the original fragment were generated. Expression of the fragments showed that they form mostly stable proteins, with evidence of them adopting a comparable secondary structure, with the proteins showing a tendency to oligomerize in solution.

Results and Discussion

The fragment

The fragment that has been identified within the RBP has been proposed to share an evolutionary relationship between the originating Periplasmic-binding protein (PBP) like fold and other folds (Ferruz et al., 2021). The fact that not just structural but also sequence evidence for a possible evolutionary relationship can be found supports the hypothesis that this *fragment* originated from a common progenitor. Because of its distribution between various folds in vastly different contexts, it seems likely that its origins lie early in the evolution of proteins. Structurally, the *fragment* consists of the first $\alpha_3\beta_4$-element of the RBP (Figure 1A). Based on the crystal structure for the full-length protein (PDB-ID: 2FN9), and assuming it will keep that structure, the *fragment* would also consist of the central beta-sheet, with the three α-helices flanking it on each side (Figure 1B).

Since binding in PBPs is usually facilitated by an interface between two lobes and thus distributing the interacting residues over the entirety of the sequence, the contribution of the *fragment* to binding of the canonical ligand ribose is limited. Only two residues – the two asparagine at position 14 and 65 – are present in the *fragment*, making a binding of ribose of just the *fragment* highly unlikely.

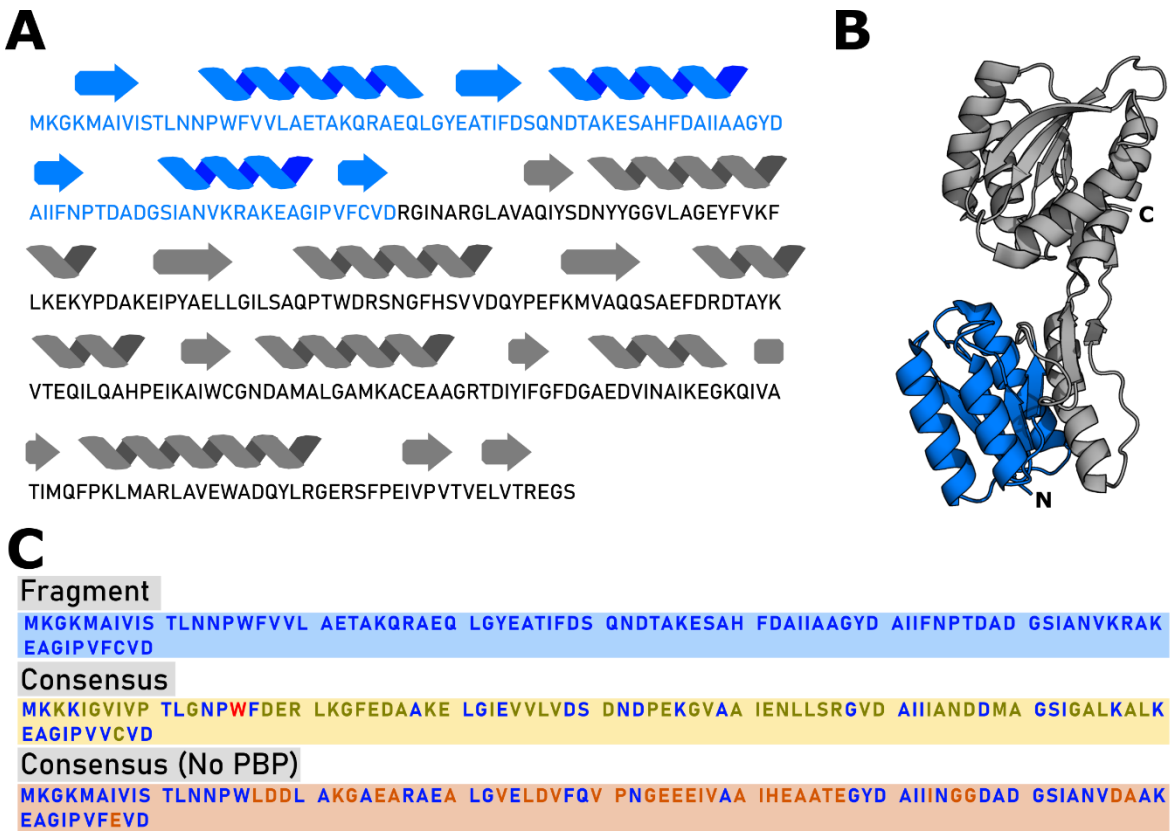


Figure 1. Sequence and structural context of the fragment within the parental protein and sequence of the consensus fragments. (A) Sequence of *Thermotoga maritima* RBP with secondary structure elements (transferred from the PDB entry 2FN8) colored in grey and blue for the fragment. (B) Cartoon representation of RBP (2FN8) in grey, with the fragment highlighted in blue on the N-terminal lobe of the bilobal periplasmic binding protein fold. (C) Sequences of the fragment, cFragment and cFragment$_{noPBP}$ in blue, yellow and orange respectively. Changes according to the consensus sequence for each are highlighted in the respective color. The fixed position for the tryptophane at position 15 for cFragment is highlighted in red.

Changing the fragment – the consensus sequence

To probe the independence of the *fragment* from the structural context of its parental fold, two consensus sequences were obtained. By utilizing the multiple sequence alignment taken from HHpred, the consensus sequences including all sequences (*cFragment*) and from sequences restricted to the PBP-like fold (*cFragment*$_{noPBP}$) were generated. While the high probability cut-off of 90% limited the number of sequences included in the compiling of the consensus sequences, 204 for *cFragment* and 24 for *cFragment*$_{noPBP}$, it still induced significant changes. 52 of the 90 positions in *cFragment* and 34 in *cFragment*$_{noPBP}$ were

changed according to the two consensus sequences (Figure 1C). The sequences not belonging to the SCOP fold of PBP-like I belonged to different superfamilies of either flavodoxin-like folds, the MurCD N-terminal domain or the Chelatase-like and Phosphofructokinase fold. However, the relationship between the PBP-like fold and flavodoxins has been described before, explaining the main influence of flavodoxins on the $cFragment_{noPBP}$ (Ferruz, 2021).

Most changes in the sequence of both consensus fragments are observed in the $\alpha\beta_2$-element but found throughout the entirety of the structure. Due to the nature of the substitutions, the changes were introduced in one step, making it difficult to predict changes in the stability and structure of the different constructs.

The fragment generated from RBP is soluble and has a defined secondary structure

To investigate the influence of isolating the *fragment* on its structural makeup, its structure was first analyzed spectroscopically. To determine the folding state and size of the *fragment* the protein was characterized using circular dichroism (CD), its intrinsic fluorescence (IF) and multi-angle light scattering (MALS). Far-UV CD spectra for the *fragment* show minima at 222 and 214 nm (Figure 2A), consistent with the expected α/β-layer secondary structure. Additionally, it corresponds well with the spectra obtained for the parental protein, indicating that the secondary structure elements of the *fragment* are formed in a comparable way. Similarly, the fluorescence spectrum of the *fragment* shows a maximum at a wavelength of 340 nm, which is close to that of the full-length RBP at 336 nm (Figure 2B), indicating that the single Tryptophane – although its exposed position – still in a polar environment. This could be a hint for the correct formation of the expected tertiary structure, although it is impossible to tell from only a single aromatic residue. However, the protein is not completely unfolded. This is further corroborated by the MALS analysis, which shows

two peaks of defined molar mass. At a protein concentration of 1 mg ml^{-1} of the *fragment* the first and major peak shows a molar mass of 15.5 kDa, which doesn't correspond to the expected mass of around 10.6 kDa (Table 1). There also seems to be a concentration dependent shift of this major peak to higher molecular weight with increasing protein concentration, shifting from 15.5 kDa at 1 mg ml^{-1} to 19.1 kDa at a protein concentration of 5 mg ml^{-1}. Additionally, there is a second peak at a mass of 30.6 kDa which does not show the same concentration dependent shift in molecular mass but is directly proportional to the protein signal. A possible explanation for this behavior could be the formation of a dynamic equilibrium of monomer to trimer. This is also supported by the fact that all the peaks are well resolved, and do not change their signal intensity or retention profile in a time-course measurement (data not shown), indicating it being entirely concentration dependent.

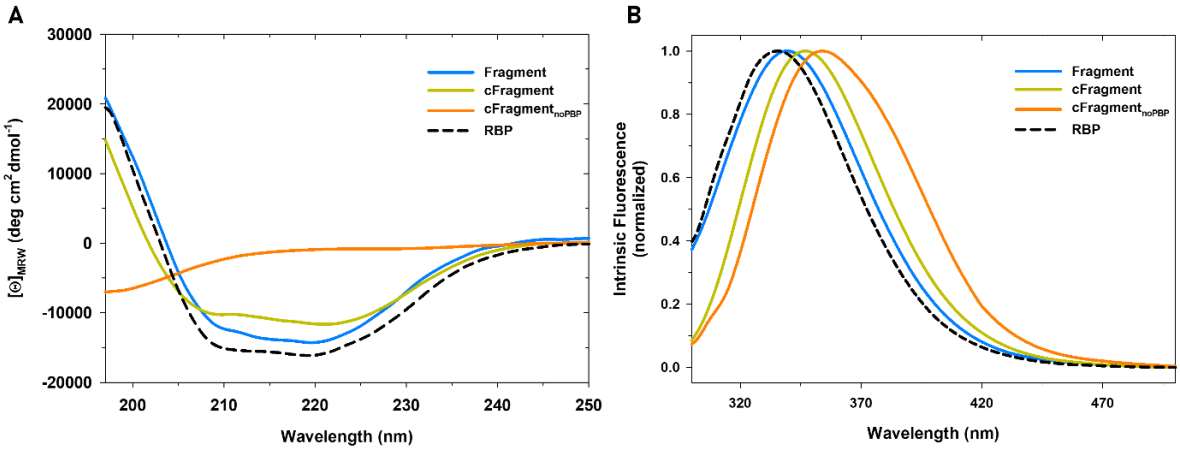


Figure 2. Analysis of the structure of the Fragments in comparison to RBP. Spectra of Fragment (blue), cFragment (yellow) and cFragment$_{noPBP}$ (orange) in solid lines and RBP (black) in dashed lines for both Far-UV-CD (A) and intrinsic fluorescence (B)

The consensus sequence shows different behavior to the original fragment

Comparing the structural characteristics of the *cFragment* and *cFragment$_{noPBP}$* to the original one and the parental protein shows significant changes in their characteristics. While the far-UV-CD of the *cFragment* still shows comparable secondary structure content,

secondary structure in the *cFragment*$_{noPBP}$ seems to have been almost completely lost, with only a small, undefined negative signal at wavelengths < 225 nm (Figure 2A). Corresponding behavior can be observed for the IF spectra, with the maximum of fluorescence at 347 nm and 354 nm for *cFragment* and *cFragment*$_{noPBP}$ respectively (Figure 2B). This shift to higher wavelengths would also correspond with the tryptophane being more exposed to solvent in the *cFragment* than in the original fragment, and almost completely so in *cFragment*$_{noPBP}$. While the *cFragment* is indicated to have a secondary and tertiary structure similar to RBP, *cFragment*$_{noPBP}$ appears to have lost all of its structural features. Despite this apparent loss of structure, the protein is still perfectly soluble, and does not aggregate at protein concentrations of 20 mg ml^{-1}.

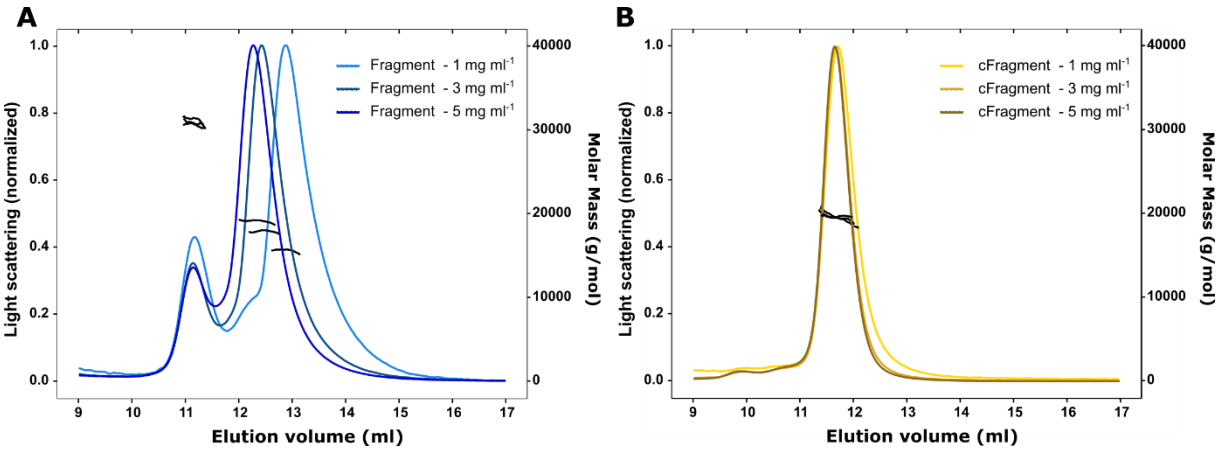


Figure 3. Determination of the molecular mass of the fragments. SEC-MALS analysis of the (A) fragment (blue) and (B) cFragment (yellow) at 1 mg ml^{-1}, 3 mg ml^{-1} and 5 mg ml^{-1}. Measurements at different concentrations are plotted as the normalized light scattering signal against the retention volume, with detected molecular volume represented by black lines in each plot.

The subsequent MALS analysis agrees with these results. While the measurement of *cFragment* at different protein concentrations resulted in a single peak at 19.6 kDa (Fig 3B), light scattering of the *cFragment*$_{noPBP}$ didn't yield any defined peaks. This indicates that while *cFragment* appears to form a stable dimer (expected mass of the monomer being 10.5 kDa, see Table 1) at all measured concentrations, *cFragment*$_{noPBP}$ does not assume a well-defined structure.

Table 1. Molecular weight determination with SEC-MALS.

Sample (concentration)	Expected Mw (kDa)	Peak I Experimental Mw (kDa)	Uncertainty (%)	Peak II Experimental Mw (kDa)	Uncertainty (%)
Fragment (1.0 mg mL^{-1})	10.7	15.5	0.4	30.6	0.8
Fragment (3.0 mg mL^{-1})		17.8	0.6	30.6	0.4
Fragment (5.0 mg mL^{-1})		19.1	0.4	31.2	0.3
cFragment (1.0 mg mL^{-1})	10.6	19.5	0.5	-	-
cFragment (3.0 mg mL^{-1})		19.6	0.7	-	-
cFragment (5.0 mg mL^{-1})		19.7	0.5	-	-

Conclusion

The idea that some fundamental components of proteins are recurring sub-domain fragments has been proposed several times in recent years, however there is still a poor understanding what the governing principles of these mechanisms are. Are these fragments just remnants of a long-lost function, or are they anchor points for the folding of the entire protein? While the first would require extensive sequence analysis, reconstruction and functional investigation, the latter could possibly be accessed through the behavior of these fragments in modern proteins. Not only is investigating these fragments in their modern structural context be an interesting concept but could also further our understanding of how proteins use a pre-defined set of building blocks.

The fragments derived from the ribose binding protein of *Thermotoga maritima* have already been described to contain many sequence and structural similarities to other superfamilies. Since this reuse of subdomain-sized fragments has predominantly only been observed *in silico*, this analysis of such a fragment *in vitro* can help shed light on why. Not only is the unchanged *fragment* taken directly from the RBP a stable protein but appears to have a comparable structure. The same applies to the consensus fragment. However, the

spectroscopic measurements also indicate a loss of overall secondary structure, and changes in the tertiary structure.

Also, the propensity of both the *fragment* and *cFragment* to form stable and defined oligomers in solution could be evidence of them tending to form protein-protein interfaces. This modularity coupled with their robustness could have been a major contributing factor of their successful propagation during evolution. To generate a clearer case for the argument of the structural importance of these elements rather than their contribution to function the investigation of the atomistic structure and folding studies would need to be conducted with these fragments. It is also unclear where the lack of structure in $cFragment_{noPBP}$ originates from. To generate a better overview of whether fragments can be easily taken from their structural context, different fragments from a variety of folds should be isolated and investigated as well. This way not just a repertoire of robust building blocks could be compiled, but understanding the principles behind this propagation throughout the protein structural universe could help us understand better how proteins fold and continue to evolve.

Materials and Methods

Identification of the protein fragments and sequence analysis

The previously described N-terminal fragment in *Thermotoga maritima* RBP (Ferruz et al. 2021) was used as a basis for the generation of the consensus sequences. To obtain the consensus sequences, the first 90 residues of RBP were used to generate a multiple-sequence alignment utilizing the HHpred program built into the MPI Bioinformatics Toolkit (Zimmermann et al. 2018) using standard parameters, but only including results of 90% HHpred probability or higher. The consensus fragment is the resulting consensus sequence of the 204 sequences found in the analysis, regardless of their protein fold. Positions where no consensus was found were kept according to their identity in the original RBP sequence.

Position 15 was deliberately kept as the original tryptophane, even though consensus suggested a phenylalanine at the position to still have access to spectroscopic methods. The Consensus Fragment excluding sequences from the same PBP-like fold (Consensus No PBP) was calculated analogously, however excluding all but the 24 sequences found not to be of the same fold (Figure 1C).

Cloning and generation of RBP-constructs

Gene synthesis and cloning for the different fragments was done by Biocat, all carrying an additional N-terminal His_6-tag. The constructs of the fragment, consensus, and consensus (No PBP) was cloned into pET21-vectors. Individual clones were obtained by transforming *Escherichia coli* BL21 (DE3) cells (Merck Millipore Novagen) by adding 50 ng of purified plasmid, heat shock and subsequent plating on agar-plates supplemented with 100 µg mL^{-1} ampicillin. The parental RBP was purified as described in [anderes paper]

Expression and purification of fragment constructs

The transformant *E. coli* BL21(DE3) were grown in *Terrific broth* media (TB) at 37 °C to an OD_{600} of 1.2 in the presence of 100 µg mL^{-1} ampicillin. Protein expression was induced by the addition of Isopropyl-β-thiogalactopyranoside to a concentration of 1 mM and a total time of 18 h at 20 °C. Cells were harvested via centrifugation (5000 × G, 15 min), resuspended in binding buffer (20 mL g^{-1} wet weight), lysed by sonication, and subsequently centrifuged to remove remaining cell debris (40000 × G, 1 h). The cleared lysate was filtered through a 0.22 µm filter previous to the affinity column step.

Subsequent Immobilized Metal Ion Chromatography (IMAC) was performed on a Cytiva HisTrap 5 mL column previously equilibrated with buffer binding (20 mM sodium phosphate, 500 mM sodium chloride, 10 mM imidazole, pH 7.8). Elution was performed with a step of IMAC-Elution-Buffer (20 mM sodium phosphate, 500 mM sodium chloride,

600 mM imidazole, pH 7.8) at 40%, and fractions corresponding to the eluted protein pooled and concentrated to a volume suitable for the size exclusion chromatography step.

Size exclusion chromatography was performed as final purification step for all constructs on a Cytiva Superdex 26/600 75 pg with an isocratic elution using buffer 20 mM sodium phosphate, 50 mM sodium chloride, pH 7.8. Fractions consistent with the proteins of interest were analyzed by SDS-PAGE, pooled, flash frozen in liquid nitrogen, and stored at -20°C until further analysis.

Far-UV Circular Dichroism (CD)

Far-UV Circular Dichroism (CD) measurements were performed at 20 °C in buffer 20 mM sodium phosphate, 50 mM sodium chloride, pH 7.8 in a Jasco J-710 spectropolarimeter equipped with a Peltier device to control temperature (PTC-348 WI). Spectra were collected using 10 µM protein concentration in a 2 mm cuvette, 195-250 nm wavelength range, and 1 nm bandwidth. After buffer subtraction, raw data were converted to mean residue molar ellipticity ([Θ]) with $[\Theta] = \Theta / l\ C\ N_r$, where Θ is the ellipticity signal in millidegrees, l is the cell path in mm, C is the molar protein concentration, and N_r is the number of amino acids per protein (Greenfield, 2007).

Intrinsic Fluorescence (IF)

Intrinsic fluorescence (IF) spectra were collected on a Jasco FP-6500 spectrofluorometer coupled with a water bath (Julabo MB) to control the temperature. Experiments were performed at 20 °C in buffer 20 mM sodium phosphate, 50 mM sodium chloride, pH 7.8 and 10 µM protein concentration, with 280 nm as excitation wavelength, 300-500 nm as emission wavelength range, and 1 nm bandwidth. Raw signal was normalized for total signal strength.

Analytical Size Exclusion Chromatography coupled with Multi Angle Light Scattering (SEC-MALS)

Analytical Size Exclusion Chromatography measurements were performed coupled to a miniDAWN Multi Angle Light Scattering (MALS) detector and an Optilab refractometer (Wyatt Technology). Samples previously centrifuged and filtered were run in a Superdex 75 Increase 10/300 GL column connected to an Äkta Pure System (GE Healthcare Life Sciences) equilibrated with buffer 20 mM sodium phosphate, 50 mM sodium chloride, 0.02% sodium azide, pH 7.8. Experiments were collected at room temperature with a protein concentration of 1.0 mg mL^{-1}, 3.0 mg mL^{-1} and 5.0 mg mL^{-1} at a 0.8 mL min^{-1} flow rate. Reproducibility during all SEC-MALS collections was tested by running a BSA standard sample at 2 mg mL^{-1} at the beginning and end of all experiments, resulting in identical results. Collection of the experiments and data analysis were done using ASTRA v.7.3.2 software (Wyatt Technology).

Acknowledgments

We thank Sabrina Wischt and Sooruban Shanmugaratnam for their competent technical support. We acknowledge all the members of Höcker Lab for their constructive suggestions to improve the research.

Funding

This work was supported by the European Research Council (ERC Consolidator Grant 647548 ‘Protein Lego’ to B.H.)

Competing interests

The authors declare that they have no conflicts of interest with the contents of this article.

Author Contributions

F.M. and B.H. designed the research, F.M., T.K. purified the different constructs, T.K., F.M. collected and analyzed CD, IF, and SEC-MALS data, F.M., T.K. performed the DSC experiments. F.M., B.H. wrote the manuscript. All authors discussed and commented on the manuscript.

Data and materials availability

All data to support the conclusions of this manuscript are included in the main text and supplementary materials.

References

1. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic acids research*, 42(D1), D310-D314.
2. Ian Sillitoe, Natalie Dawson, Tony E Lewis, Sayoni Das, Jonathan G Lees, Paul Ashford, Adeyelu Tolulope, Harry M Scholes, Ilya Senatorov, Andra Bujan, Fatima Ceballos Rodriguez-Conde, Benjamin Dowling, Janet Thornton, Christine A Orengo, CATH: expanding the horizons of structure-based functional annotations for genome sequences, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D280–D284, https://doi.org/10.1093/nar/gky1097
3. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. (2014) ECOD: An Evolutionary Classification of Protein Domains. PLoS Comput Biol 10(12): e1003926. https://doi.org/10.1371/journal.pcbi.1003926
4. Farías-Rico, J., Schmidt, S. & Höcker, B (2014). Evolutionary relationship of two ancient protein superfolds. *Nat Chem Biol* **10,** 710–715. https://doi.org/10.1038/nchembio.1579
5. Alva, V. Söding, J., Lupas A.N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. eLife, 4, Article e09410. https://doi.org/10.7554/elife.09410
6. MacKenzie C.O., Zhou J., Grigoryan G. (2017). Tertiary alphabet for the observable protein structural universe. Proc Natl Acad Sci USA, 113, 7438-7447. https://doi.org/10.1073/pnas.1607178113
7. Nepomnyachiy S., Ben-Tal N., Kolodny R. (2017). Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. Proc Natl Acad Sci, 114, 11703-11708. https://doi.org/10.1073/pnas.1707642114
8. Ferruz N., Lobos F., Lemm D., Toledo-Patino S., Farías-Rico J.A., Schmidt S., Höcker B. (2020). Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design. *Current Opinion in Structural Biology*, 13, 3898-3914. https://doi.org/10.1016/j.jmb.2020.04.013
9. Konagurthu AS, Subramanian R, Allison L, Abramson D, Stuckey PJ, Garcia de la Banda M and Lesk AM (2021) Universal Architectural Concepts Underlying Protein Folding Patterns. Front. Mol. Biosci. 7:612920. doi: 10.3389/fmolb.2020.612920
10. Alva V., Remmert M., Biegert A., Lupas A.N., Söding J. (2009) A galaxy of folds. Protein Science, 19, 124-130. https://doi.org/10.1002/pro.297
11. Ohta, T. (2000). Mechanisms of molecular evolution. Phil. Trans. R. Soc. Lond., B355, 1623-1626. http://doi.org/10.1098/rstb.2000.0724

8. Paper V

Romero-Romero S., Kordes S., **Michel F.**, Höcker B.

Evolution, folding, and design of TIM barrels and related proteins.

Current Opinion in Structural Biology, 2021, 68, pp. 94-104

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Evolution, folding, and design of TIM barrels and related proteins

Sergio Romero-Romero[1], Sina Kordes[1], Florian Michel[1] and Birte Höcker

Proteins are chief actors in life that perform a myriad of exquisite functions. This diversity has been enabled through the evolution and diversification of protein folds. Analysis of sequences and structures strongly suggest that numerous protein pieces have been reused as building blocks and propagated to many modern folds. This information can be traced to understand how the protein world has diversified. In this review, we discuss the latest advances in the analysis of protein evolutionary units, and we use as a model system one of the most abundant and versatile topologies, the TIM-barrel fold, to highlight the existing common principles that interconnect protein evolution, structure, folding, function, and design.

Address

Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany

Corresponding author: Höcker, Birte (birte.hoecker@uni-bayreuth.de)
[1] Equal contribution.

Current Opinion in Structural Biology 2021, **68**:94–104

This review comes from a themed issue on **Sequences and topology**

Edited by **Nir Ben-Tal** and **Andrei N Lupas**

For a complete overview see the Issue and the Editorial

Available online 13th January 2021

https://doi.org/10.1016/j.sbi.2020.12.007

Introduction

Structural and functional diversity in modern proteins is the result of diversification and optimization processes over the course of evolution. Studying these processes is useful to evaluate how different molecular mechanisms, like duplication and recombination, shape biophysical properties in proteins. Sequence and structural analysis suggest that numerous protein pieces, considered as evolutionary units, have been reused and combined to create higher complexity. In this context, what are the reasons for the recurring success of some of these units? What is their role in protein fold diversification? And how can we use the accumulated information to further our protein design goals?

In this review, we try to unravel these mysteries by integrating different perspectives and approaches (Figure 1). We first discuss the current views of evolutionary units (Section 'Current views of evolutionary units'). Then, we use the TIM-barrel fold as model system to analyze how our knowledge of the protein-based world is enhanced by the integration of evolutionary analysis (Section 'Evolutionary events: fragments and natural TIM-barrel proteins'), experimental recreation of evolutionary events (Section 'Recreating evolutionary events in the lab: chimeragenesis and directed evolution'), folding-function-fitness studies (Section 'Three *f* determinants in TIM-barrel evolution: *f*olding, *f*unction, and *f*itness'), and protein design approaches (Section 'Learning from nature towards protein design'). We illustrate how these studies pave the way to a detailed description of existing structure-folding-function-fitness relationships and also boost the design of new proteins with novel molecular properties.

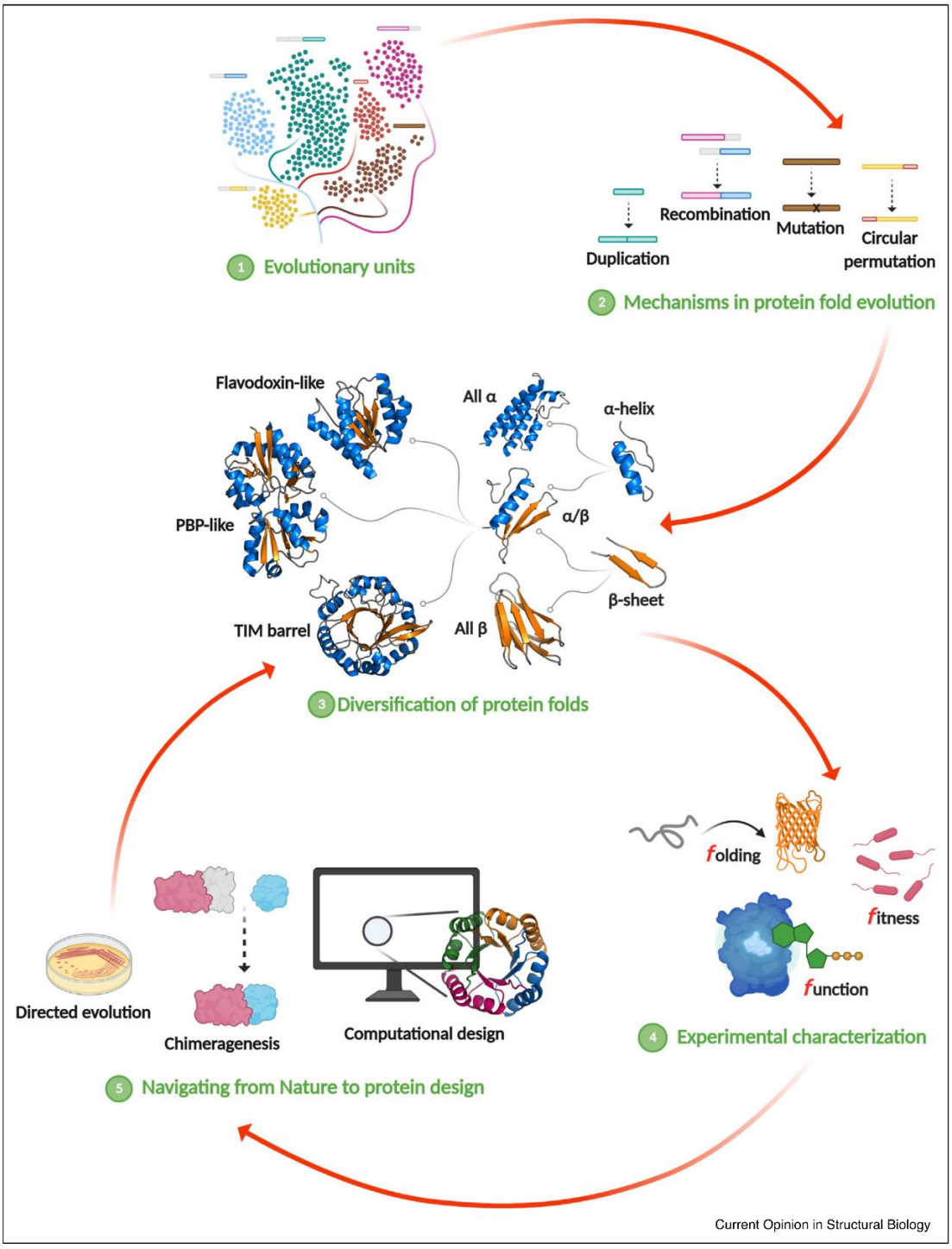
Current views of evolutionary units

Look at any protein and you are bound to find pieces that appear to have been reused either in different proteins or as the modules in a repeat protein. Clearly, reuse of sequences is ubiquitous within the natural fold space as was suggested already early on [1,2]. For protein scientists this beckons the question: how many of these pieces are there and what makes them so successful?

The structural annotation of proteins typically includes consulting at least one of the major databases SCOP, CATH or ECOD [3–5] to append additional information on evolutionary relationships. Molecular evolution studies have shown that different forces and mechanisms such as mutations, duplications, recombinations, deletions, and circular permutations drive the diversification of the protein-based world [6,7]. These mechanisms also hold true for events in the subdomain regime.

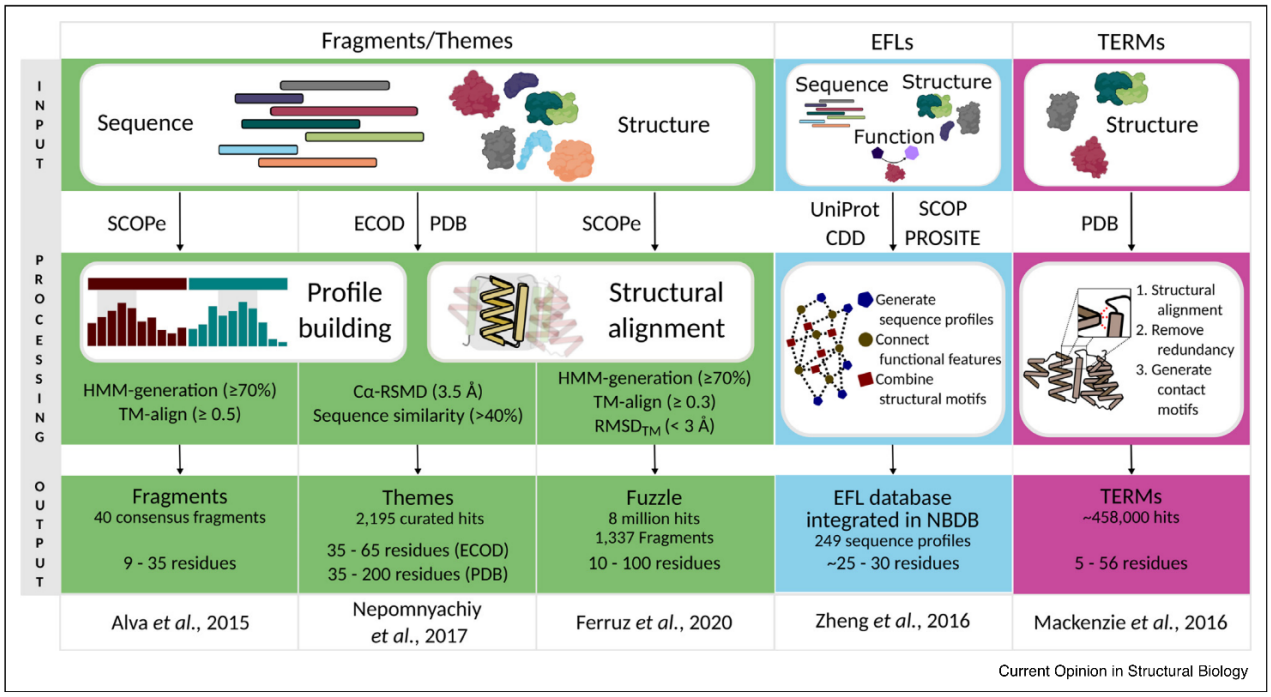
In recent years there have been several approaches to define subdomain units as distinguishable building blocks (Figure 2). For example, an evolutionary relationship between the TIM-barrel and flavodoxin-like folds based on a 40-residue fragment was identified by sequence searches [8]. In a large-scale approach, Alva *et al.* identified and defined the reuse of elements within all modern proteins [9]. They generated a vocabulary of 40 subdomain fragments of up to 38 residues, which occur within a

Figure 1



Schematic overview of the relationships between protein fold evolution, experimental characterization, and design approaches discussed in this review. The upper part of the figure shows how evolutionary units are reused through different molecular mechanisms to diversify protein folds. Experimental reconstruction of different evolutionary pathways and the analysis of folding, function, and fitness determinants in evolution increase our knowledge of the protein-based world and allow navigating from Nature to protein design as shown in the bottom part.

Figure 2



Current subdomain classification approaches. Shown is the generation of available subdomain databases including the different input, data processing, and final output. While Fragment/Themes are continuous sequences and are defined by HMM-profile comparisons and structural alignments, TERMs are non-continuous and focus on contact maps for classification. In contrast, EFLs combine information from structure, sequence and function, but are limited by existing annotation of functional sites.

great number of different folds. Subsequent efforts to expand on these initial fragments led to the description of *themes* – reused fragments of at least 35 residues [10••]. A *theme* is defined whenever a sensitive sequence search using HHsearch suggests remote homology.

Along the same lines, Ferruz *et al.* expanded the fragment universe applying a set of filters to ensure the fragments are related, but not restricting their length [11••]. This generated a dataset of over eight million hits, which are summarized in the *Fuzzle* database (https://fuzzle.uni-bayreuth.de). When visualizing the dataset in a network representation a major component is observed that includes many hits between folds thought to be ancestral reinforcing earlier observations on different datasets [12,13]. This might hint not only to a common evolutionary history, but also to the existence of a favorable set of rules for protein folding, function, and fitness.

Another description by Berezovsky defines *elementary functional loops* (EFLs) [14]. These EFLs describe stretches of proteins with a specific sequence profile thought to be defined by the polymer nature of the polypeptide as reviewed recently [15]. Combining this with information on the conservation of structure and function provides indications, which elements might have proven successful in a primordial peptide-stage of evolution. This concept has been employed for example in the *nucleotide binding database* (NBDB), which contains EFLs involved in binding nucleotide-containing ligands [16]. Phosphate binding signatures obtained by this database were applied in the design of a P-loop protein testing the role of polymer physics in the emergence of basic units of proteins [17].

A fourth view that does not necessarily focus on the evolutionary aspect but rather on protein fold space are the *tertiary structural motifs* (TERMs) [18]. TERMs are 5–56 residue-long, discontinuous structural entities that are generated solely by comparing their environment. While TERMs focus primarily on conserved structural environments, a comparison of motifs generated by simulated evolution on TERMs and those of their natural counterparts showed that TERMs were able to accurately describe nature-like sequence variation.

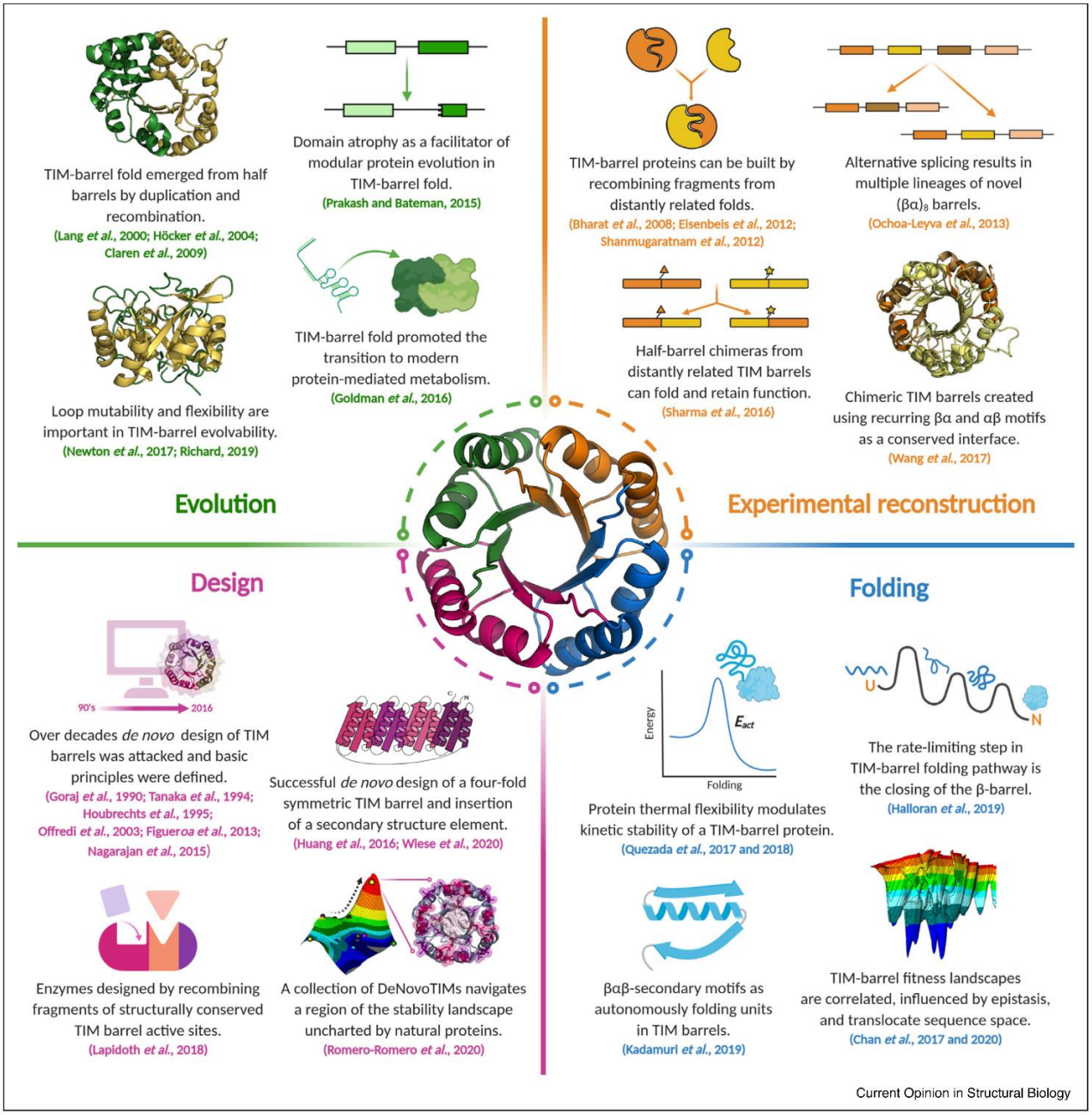
These examples of either using structural information alone or sensitive in-depth sequence analysis or a combination thereof clearly hint to one thing: there is a subset of

successful ancestral sequences that are to this day propagated to many modern folds.

Evolutionary events: fragments and natural TIM-barrel proteins

The previous section showed that, even after a considerable timespan, we can detect evolutionary relationships in modern proteins. Can we decode the underlying mechanisms of conservation of subdomain fragments in natural proteins? This general question has been explored by analyzing the evolution of different protein folds, for which the TIM barrel is a model system (Figure 3). This fold is regarded to be one of the oldest and encompasses a wide variety of known protein functions [19–21]. Its

Figure 3



Summary of recent central studies that interconnect the evolution, its experimental reconstruction, folding, and design of TIM-barrel proteins as discussed in this review.

canonical fold consists of a central eight-stranded, parallel β-barrel surrounded by eight α-helices forming the eponymous $(\beta\alpha)_8$-barrel structure. It has previously been shown that subdomain parts of the TIM-barrel fold present an excellent model to probe the role of subdomain events, but also explore its evolution [22].

In a recent endeavor, Kadamuri *et al.* theorized that a set of βαβ sequences exists within the TIM-barrel fold-space, which would be autonomously folding units [23]. While there are not yet any reports of natural βαβ motifs folding in isolation, investigating the subdomain folding regime in TIM barrels might reveal crucial steps to improve the creation of novel proteins and help elucidate the evolution of protein domains themselves.

A study by Michalska *et al.* on the structural flexibility of naturally occurring TIM barrels reported a 3D-domain swap of an $(\alpha\beta)_2$ element within a tryptophan synthase structure [24]. A similar event has recently been observed in a crystal structure of the archaeal chemotaxis protein CheY [25]. An analysis of alternative splicing events of $(\beta\alpha)_8$ barrels within the human genome also showed a considerable fraction expressing only as subdomains, and are thought to assemble to a complete barrel with their complementary partners [26]. These observations hint at a flexible subdomain composition within α/β proteins. This concept has been experimentally explored as will be discussed further in Section 'Recreating evolutionary events in the lab: chimeragenesis and directed evolution'.

When Prakash and Bateman analyzed the variation of TIM-barrel domain boundaries, they found what they propose to be *domain atrophy* [27]. This rare event is characterized by a loss of core secondary structure features that is potentially detrimental to domain stability. While it is still not clear why such events are evolutionary fixed, a possible rescue of stability appears to be the formation of protein-protein interactions, for example, in homodimers.

All these examples of subdomain evolutionary events in the TIM-barrel fold point to one thing: there is a propensity of some proteins to swap subdomain elements. To really gauge if this subdomain recombination played — or still plays — an important role in the diversification of proteins, more protein folds need to be examined. Understanding the common principles that govern this process could help improve our knowledge of protein stability, folding, function and evolution.

Recreating evolutionary events in the lab: chimeragenesis and directed evolution

The enormous diversity of protein structures and functions can be interpreted as the result of a massive *experiment* that has been carried out by Nature in a sustained way for millions of years, whose results are observed in the broad number of protein sequences and structures. In the previous section, we discussed that diverse evolutionary events in natural proteins allow the expansion of the protein fold space. Now, we focus on how some of these evolutionary events can be recreated in the laboratory through chimeragenesis and directed evolution. Both approaches offer a good alternative to test evolutionary and thermodynamic hypotheses and also to generate novel proteins (Figure 3).

Newton *et al.* explored the evolution of the TIM-barrel enzyme HisA using directed evolution techniques [28••]. They follow up on the innovation-amplification-divergence model previously proposed as an explanation of how gene duplication leads to proteins with new functions [29]. They show how beneficial substitutions selected during real-time evolution can result in manifold changes in enzyme function and bacterial fitness. The results emphasize the importance of loop mutability and confirms the TIM barrel as an inherently evolvable protein scaffold.

The current evolutionary hypothesis about the emergence of the TIM-barrel fold is that it evolved from duplication and fusion events of a half barrel, that is, a $(\beta\alpha)_4$ or even smaller units [30–33]. This possible pathway has been tested computationally and experimentally by analyzing sequence, structural, and folding properties [32,34,35]. Following this idea, Sharma *et al.* engineered and characterized active and stable chimeric TIM barrels of two distantly related glycosyl hydrolases, demonstrating that half-barrel domains from different sources can assemble and adopt the pre-evolved function [36]. Likewise, Almeida *et al.* tested the idea that $(\beta\alpha)_4$ halves are self-contained evolutionary units, independent of their size and internal symmetry. They introduced mutations in the inter-half contacts of a β-glucosidase to obtain independent half barrels that unfold cooperatively [37]. Further, Wang *et al.* identified physicochemical properties from a set of non-redundant TIM-barrel proteins that strongly support the existence of recurring βα and αβ motifs in this fold [38]. In addition, using a conserved αβα element as a recombination site, they created a chimeric protein from two different TIM barrels, highlighting the potential of recurring motifs as naturally optimized interfaces to engineer well-folded chimeras.

Inspired by TIM-barrel modularity, Lapidoth *et al.* designed highly active and stable enzymes by creating fragments of structurally conserved sites of two unrelated TIM-barrel families and then assembled them to create a large set of combinatorial backbones [39••]. The reported computational approach mimics natural evolutionary processes such as recombinations, insertions, deletions, and mutations, but it is more radical than these individual events since all of them are applied simultaneously to modify the protein fitness.

As will be discussed in the last section (Learning from nature towards protein design), this method could be extended to create new biocatalysts by combining more distantly related families.

Apart from recombination events within a protein fold, recombination of heterologous structural motifs of unrelated folds is possible. Although difficult to detect in Nature, the idea can be tested in the laboratory and might be used to design proteins with novel biophysical properties [21]. In this context, ElGamacy *et al.* engineered an asymmetric dRP lyase fold fusing two heterologous and unrelated supersecondary structures. After interface optimization the approach generated a stable chimera with high precision to the original design [40•].

Similarly, we have used chimeragenesis in the past to elucidate evolutionary relationships of several α/β folds and design new proteins. Chimeras built combining parts of the flavodoxin-like proteins CheY or NarL with a piece of the TIM barrel HisF demonstrate that $(\beta\alpha)_8$-barrel proteins can be constructed by recombining a large repertoire of natural protein fragments from distantly related folds [8,41–43]. This interchangeability offers a great opportunity to retrace early evolutionary steps. Following up on this, Toledo-Patiño *et al.* found sequence-based evidence that the singleton HemD-like fold emerged from the flavodoxin-like fold [44•]. To test the hypothesized path, consisting of insert-assisted segment swap, gene duplication, and fusion, these evolutionary events were experimentally reverted, yielding well-folded and stable proteins. The results strongly support the emergence of the HemD-like fold from flavodoxin-like proteins and highlight the importance of duplication and fusion as evolutionary events that allow the creation of complex proteins. These experimental reconstructions of possible evolutionary events fit well with the bioinformatic studies on protein fragments as discussed in section 'Current views of evolutionary units'. Databases such as *Fuzzle* [11••] provide many starting points for similar evolutionary explorations and open new ways to use already existing sequences in protein design. Fragments identified in *Fuzzle* can be used directly in the tool *Protlego* (https://hoecker-lab.github.io/protlego/) for automated chimera design and analysis [45].

Three *f* determinants in TIM-barrel evolution: *folding*, *function*, and *fitness*

The evolutionary study of biophysical determinants is useful to evaluate the role evolution has on the physical properties of proteins and informs us on how changes in the amino acid sequence shaped function in a specific fold [46]. In this section, we focus on recent advances to understand the biophysical basis underlying the success of the TIM-barrel fold as one of the most robust and versatile scaffolds.

The TIM-barrel fold provides a good architecture to explore how folding mechanisms have been conserved or diverged during evolution (Figure 3). In this context, Halloran *et al.* analyzed on a molecular level the earliest events in the folding of a TIM-barrel protein [47••]. Experimental and computational approaches revealed that the kinetic intermediate commonly observed in TIM barrels is dominated by a native-like structure in the central region of the sequence. They determined the rate-limiting step in the folding pathway to be the frustration encountered by the competition between the N-terminus or C-terminus to close the internal β-barrel. Also analyzing TIM-barrel proteins, Romero-Romero *et al.* studied and compared the folding pathway of eukaryotic homologous triosephosphate isomerases. Structural and biophysical analysis suggested that interfacial water molecules and water-mediated interactions could modulate the number of equilibrium intermediates, and therefore, the folding pathway in this enzyme family [48].

TIM-barrel proteins are notable for their diversity in catalytic activities. The broad presence of this topology in different enzymes has led to the assumption that the TIM-barrel fold played a central role in early evolution of catalysis. In a bioinformatic study, Goldman *et al.* showed by comparing the functional diversity of different protein folds that TIM-barrel proteins use the broadest range of enzymatic cofactors, including some putatively ancient cofactors [49•]. This supports the idea that the TIM barrel represented an ideal scaffold to facilitate the transition from ribozymes, peptides, and abiotic catalysts to modern protein-mediated metabolism.

Likewise, in terms of protein flexibility and enzymatic catalysis, Richard recently discussed why the selection and optimization of protein folds with multiple flexible loops, such as the TIM-barrel topology, is favored during enzyme evolution [50•]. He proposes that in TIM barrels the exploration of many different conformations during loop movement provides a potential starting point for the evolution of a new enzyme activity and allows the conformational changes needed in floppy enzymes. Also related with protein flexibility, but in the context of stability and evolution, Quezada *et al.* analyzed the molecular basis of the kinetic stability differences of two related triosephosphate isomerases and engineered new functional TIM-barrel enzymes with fine-tuned stabilities [51,52•]. They found a correlation between thermal flexibility and kinetic stability, suggesting how evolution has reached a balance between function and stability in cell-relevant timescales.

The evolution of protein folding, function, and fitness can be seen as a walk through sequence space, in the same way as was described 50 years ago by evolutionary biologist John Maynard Smith in his seminal work about natural selection and the concept of protein space [53].

Generally, each of these steps can be evaluated in terms of protein fitness, a measure of the effect that a property produces on the overall fitness of an organism. Following this logic, in two subsequent works the Matthews lab performed a quantitative description of the fitness landscape of distant orthologous TIM-barrel proteins to understand their evolutionary dynamics [54••,55]. They detected that the fitness landscapes are correlated and influenced by long-range epistatic interactions, and that these landscapes can be translocated in sequence space as a result of TIM-barrel fold plasticity.

The three *f* determinants in evolution discussed in this section have also been analyzed in other protein folds. Examples from the last years include discussions between the Makhatadze and Sanchez-Ruiz labs about the evolutionary validity of the minimal frustration hypothesis through the experimental characterization of ancestrally reconstructed proteins and extant homologous members of the thioredoxin family [56–58]. Also involving α/β proteins, Kukic *et al.* explored how the folding rates of Procarboxypeptidase A2 can be modulated during evolution by modification of the so-called nucleation-condensation mechanism [59]. Moreover, the Marqusee lab has made a substantial effort to understand how evolutionary pressures modify folding landscapes and tune kinetic and thermodynamic stability by characterizing one of the oldest protein folds, the RNase H-like superfamily [60–63]. Other interesting works are the analysis of the influence of folding energies on the fitness of β-lactamases [64], the study of protein folding and fitness landscapes of amidases [65], the analysis of cotranslational folding and fitness of an integral membrane protein [66], and the evolutionary history of myoglobins [67]. The information obtained both on TIM barrels and other folds has revealed unanticipated details in protein molecular evolution thereby increasing our understanding of sequence-folding-fitness relationships, which has also relevant implications for protein design.

Learning from nature towards protein design

In the previous sections we discussed the evolution of protein folds from smaller units and provided examples recreating such evolutionary events with respect to folding, function, and fitness. Same as protein engineering has been used to test evolutionary hypotheses, the gained knowledge can also be used to design new proteins. Initial protein design strategies were mostly based on parametrization of well understood folds or supersecondary structures. But in the last decades many powerful algorithms were developed to predict protein structures and design new proteins as has been recently reviewed [68].

One of the most widely used design software, namely Rosetta, uses 3-residue and 9-residue long fragments from known protein structures to sample the backbone in *ab initio* predictions [69,70]. Those fragments are a lot smaller than the previously described evolutionary units [9,10••,11••,14], however, they still can carry information about possible conformations. Additionally, some algorithms use evolutionary mechanisms as inspiration. The SEWING algorithm for instance incorporates current understanding of protein evolution, the emergence of proteins by recombination and duplication of smaller fragments: sets of structures meeting predefined requirements are generated by recombination of small structural motifs [71]. The more recently developed program d'TERMen uses the previously described TERMs by matching them to the target design and thereby determines sequence preferences [72]. Also, the approach from Lapidoth *et al.* mentioned previously is inspired by Nature and mimics evolution during the design process [39••]. The fully automated method combines recombination, insertion, deletion, and mutation events in a non-sequential manner. Initially a predefined set of structures is partitioned and then assembled to combinatorial backbones, which are finally applied to a complete sequence redesign. During this process conserved sites and residues necessary for catalysis or folding can be excluded from the design. In contrast to other enzyme design approaches it has the advantage that no transition state has to be modelled which is computationally expensive. This method was applied to homologous TIM barrels but could possibly be extended to more distantly related proteins, thereby creating new biocatalysts. While this approach, that is based on existing structures, can diversify enzyme function, it will not create proteins from scratch.

The complete *de novo* design of proteins is a task that has been explored and progressed increasingly in recent years fueled by technical advances in structure determination, modelling, and computation. An increase in *de novo* designed proteins could be further observed after Koga *et al.* defined rules for the design of idealized protein topologies as recently reviewed [73]. The value of these design rules, that relate foldability of a tertiary structure to the connection between secondary structure elements [74], in combination with improvements in design algorithms can be traced in the design progression of *de novo* TIM barrels.

Several attempts were made to design a symmetric TIM barrel from scratch to understand what makes this protein fold so successful (Figure 3). In the early 1990s, first symmetric designs were created using statistical information about barrel geometries and amino acid frequencies from few known TIM-barrel structures [75–80]. However, those parameters were not sufficient to achieve designs with natural-like properties as all exhibited molten-globule like states. With an increasing number of TIM-barrel structures, geometric parameters were improved, and newly emerging algorithms were applied to sequence design and created all-atom models. In this

way, the Martial lab was able to improve previous designs and create natural-like proteins [81,82]. Later, the solubility of one of those designs was improved by directed evolution and the three-dimensional structure was determined: it differed from the intended TIM barrel and resembled a Rossman-like fold [83]. Using the previously described rules for idealized topologies, Nagarajan *et al.* created four-fold symmetric TIM-barrel backbones [84]. Using folding simulations, they determined hydrogen bond networks and enrichment of polar residues in the pore as important features regarding the folding pathway. Those findings were applied during iterative sequence design and resulted in soluble proteins showing cooperative unfolding transitions, though structural studies indicated a molten globule.

In the meantime, Huang *et al.* also applied the rules from Koga *et al.* to design a four-fold symmetric TIM barrel [85•]. Their approach sampled backbones with different secondary structure lengths using predefined geometric restrictions followed by iterative sequence design enforcing sidechain-backbone hydrogen bonds. A circular-permutated variant, sTIM11, was soluble expressed and the design was validated by solving its three-dimensional structure. Further analysis revealed a significantly lower conformational stability compared to natural TIM barrels. In a modular approach, a collection of stabilized variants (DeNovoTIMs) was designed by improving hydrophobic packing [86••]. Structural and folding analysis showed that epistatic effects allow navigating an unexplored region of the stability landscape of natural proteins. One of these DeNovoTIMs was already used in a successful recombination with a *de novo* designed ferredoxin protein and engineered to bind lanthanide [87]. In another recent study, Wiese *et al.* extended sTIM11 by successfully incorporating a rationally designed small α-helix into a βα loop [88]. These works are first steps towards diversifying and ultimately functionalizing *de novo* TIM barrels.

The progression in the design of a TIM barrel reflects nicely the improvements of protein design in the last 30 years. Throughout all design approaches, a symmetric topology was targeted as despite rapidly increasing computational resources the modelling of large proteins is still time-consuming. Further, this process shows how important it is to understand a protein fold in detail and to know which interactions are essential for its stabilization. In this context, it would be interesting to analyze the design from Figueroa *et al.* [82] in detail and determine why this design acquired a different fold than intended [83]. Such analysis is important to improve our understanding and find deficiencies in current protein design strategies.

Additionally, protein design opens a door not only to increase and test our knowledge about folding, function, and fitness, but also to compare the properties of *de novo* proteins with naturally occurring ones. In this way, studies have shown that *de novo* proteins exhibit more complex folding pathways than natural proteins, as indicated for one of the first *de novo* designed proteins Top7, a βα protein [89]. This differs from natural small proteins which show high cooperativity in folding and a smooth free energy surface. In addition, the study of another small *de novo* protein Di-III_14, an IF3-like protein, revealed a more complex folding pathway than initially assumed [90••]. In-depth mutational and folding analysis revealed that electrostatic and hydrophobic networks affect the energy surface of this protein. Based on those findings, it was proposed to limit the number of charged amino acids, avoid charge segregation, and use a more diverse set of nonpolar side chains in future protein designs. Overall, these studies demonstrate that as we expand our exploration into sequence space by designing *de novo* proteins, we also expand our understanding of the molecular and physicochemical determinants that shaped and still modulate the protein-based world.

Conclusion and outlook

The study of protein evolution requires the integrated analysis of protein structure and stability, as well as folding, function, and fitness of proteins. There is clear evidence that modern diverse protein folds evolved via reuse of smaller units, which have been identified and described in recent years. Evolution of protein folds from smaller units via duplication has long been described, but also recombination is explored increasingly as an important mechanism. Understanding how protein diversity could emerge via these mechanisms is essential to learn how stable and functional proteins evolved and might be designed.

The ubiquitous TIM-barrel fold has been used in several studies to investigate its evolution, folding, and design. Explorations of the fold's evolutionary history and experiments recreating evolutionary events have revealed how recombination of recurring fragments can lead to new proteins and enzymes. These studies go hand in hand with detailed analyses of protein folding and determination of fitness landscapes of TIM barrels. Moreover, this knowledge has already been applied to the design of *de novo* TIM barrels illustrating how the connection between evolution, folding, and design closes to a cycle and how analysis of designed proteins can help us understand the biophysical properties of proteins even better. Altogether, these recent studies have significantly increased our understanding of the evolution of sequence-structure-function relationships, enabling us to access new protein space through design.

Conflict of interest statement

Nothing declared.

Acknowledgements

We gratefully acknowledge financial support from the Foundations Alexander von Humboldt and Bayer Science & Education (Humboldt-Bayer Research Fellowship for Postdoctoral Researchers to S.R.R.), from the European Research Council (ERC Consolidator grant 647548 'Protein Lego'), and the Volkswagenstiftung (grant 94747). Figures were created with BioRender.com and Pymol.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Eck RV, Dayhoff MO: **Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences.** *Science* 1966, **152**:363-366.

2. Fetrow JS, Godzik A: **Function driven protein evolution. A possible proto-protein for the RNA-binding proteins**. *Pac Symp Biocomput* 1998.

3. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG: **SCOP2 prototype: a new approach to protein structure mining**. *Nucleic Acids Res* 2014, **42**:310-314.

4. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, Tolulope A, Scholes HM, Senatorov I, Bujan A *et al.*: **CATH: expanding the horizons of structure-based functional annotations for genome sequences**. *Nucleic Acids Res* 2019, **47**:D280-D284.

5. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV: **ECOD: an evolutionary classification of protein domains.** *PLoS Comput Biol* 2014, **10**:e1003926.

6. Ohta T: **Mechanisms of molecular evolution**. *Philos Trans R Soc B Biol Sci* 2000, **355**:1623-1626.

7. Sikosek T, Chan HS: **Biophysics of protein evolution and evolutionary protein biophysics**. *J R Soc Interface* 2014, **11** 20140419.

8. Farías-Rico JA, Schmidt S, Höcker B: **Evolutionary relationship of two ancient protein superfolds**. *Nat Chem Biol* 2014, **10**:710-715.

9. Alva V, Söding J, Lupas AN: **A vocabulary of ancient peptides at the origin of folded proteins.** *eLife* 2015, **4**:e09410.

10. •• Nepomnyachiy S, Ben-Tal N, Kolodny R: **Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths**. *Proc Natl Acad Sci U S A* 2017, **114**:11703-11708
In this bioinformatic analysis the reuse of segments, which are similar in sequence and structure, is described. These segments, called themes, are subdomain elements with reuse traced to the amino acid position, highlighting the impact of protein evolution.

11. •• Ferruz N, Lobos F, Lemm D, Toledo-Patino S, Farías-Rico JA, Schmidt S, Höcker B: **Identification and analysis of natural building blocks for evolution-guided fragment-based protein design.** *J Mol Biol* 2020, **432**:3898-3914
Following an all-vs-all comparison of protein domain sequences, reused protein fragments were compiled into a network. This customizable network of fragments not only carries evolutionary significance, but can also function as a starting point for protein design by recombination.

12. Alva V, Remmert M, Biegert A, Lupas AN, Söding J: **A galaxy of folds**. *Protein Sci* 2010, **19**:124-130.

13. Nepomnyachiy S, Ben-Tal N, Kolodny R: **Global view of the protein universe**. *Proc Natl Acad Sci U S A* 2014, **111**:11691-11696.

14. Berezovsky IN, Guarnera E, Zheng Z: **Basic units of protein structure, folding, and function**. *Prog Biophys Mol Biol* 2017, **128**:85-99.

15. Berezovsky IN: **Towards descriptor of elementary functions for protein design**. *Curr Opin Struct Biol* 2019, **58**:159-165.

16. Zheng Z, Goncearenco A, Berezovsky IN: **Nucleotide binding database NBDB - a collection of sequence motifs with specific protein-ligand interactions**. *Nucleic Acids Res* 2016, **44**:D301-D307.

17. Romero Romero ML, Yang F, Lin Y-R, Toth-Petroczy A, Berezovsky IN, Goncearenco A, Yang W, Wellner A, Kumar-Deshmukh F, Sharon M *et al.*: **Simple yet functional phosphate-loop proteins**. *Proc Natl Acad Sci U S A* 2018, **115**:E11943-E11950.

18. MacKenzie CO, Zhou J, Grigoryan G: **Tertiary alphabet for the observable protein structural universe**. *Proc Natl Acad Sci U S A* 2016, **113**:E7438-E7447.

19. Banner DW, Bloomer AC, Petsko GA, Phillips DC, Pogson CI, Wilson IA, Corran PH, Furth AJ, Milman JD, Offord RE *et al.*: **Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Å resolution: using amino acid sequence data.** *Nature* 1975, **255**:609-614.

20. Nagano N, Orengo CA, Thornton JM: **One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions**. *J Mol Biol* 2002, **321**:741-765.

21. Sterner R, Höcker B: **Catalytic versatility, stability, and evolution of the (βα)8-barrel enzyme fold.** *Chem Rev* 2005, **105**:4038-4055.

22. Höcker B: **Design of proteins from smaller fragments-learning from evolution.** *Curr Opin Struct Biol* 2014, **27**:56-62.

23. Kadamuri RV, Irukuvajjula SS, Vadrevu R: **βαβ super-secondary motifs: sequence, structural overview, and pursuit of potential autonomously folding βαβ sequences from (β/α)8/TIM barrels.** *Methods in Molecular Biology*. Humana Press Inc; 2019:221-236.

24. Michalska K, Kowiel M, Bigelow L, Endres M, Gilski M, Jaskolski M, Joachimiak A: **3D domain swapping in the TIM barrel of the α subunit of *Streptococcus pneumoniae* tryptophan synthase**. *Acta Crystallogr Sect D Struct Biol* 2020, **76**:166-175.

25. Paithankar KS, Enderle M, Wirthensohn DC, Miller A, Schlesner M, Pfeiffer F, Rittner A, Grininger M, Oesterhelt D: **Structure of the archaeal chemotaxis protein CheY in a domain-swapped dimeric conformation**. *Acta Crystallogr Sect F Struct Biol Commun* 2019, **75**:576-585.

26. Ochoa-Leyva A, Montero-Morán G, Saab-Rincón G, Brieba LG, Soberón X: **Alternative splice variants in TIM barrel proteins from human genome correlate with the structural and evolutionary modularity of this versatile protein fold**. *PLoS One* 2013, **8**:e70582.

27. Prakash A, Bateman A: **Domain atrophy creates rare cases of functional partial protein domains**. *Genome Biol* 2015, **16**:88.

28. •• Newton MS, Guo X, Söderholm A, Näsvall J, Lundström P, Andersson DI, Selmer M, Patrick WM: **Structural and functional innovations in the real-time evolution of new (βα)8 barrel enzymes**. *Proc Natl Acad Sci U S A* 2017, **114**:4727-4732
In this work, a real-time evolution analysis is performed to understand how new TIM-barrel enzymes lead to phenotype and organismal fitness changes. The study details the structural and functional innovations of the HisA enzyme towards generalist or specialist activities providing clues about evolution from atomic to whole-organism levels.

29. Näsvall J, Sun L, Roth JR, Andersson DI: **Real-time evolution of new genes by innovation, amplification, and divergence.** *Science* 2012, **338**:384-387.

30. Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M: **Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion.** *Science* 2000, **289**:1546-1550.

31. Gerlt JA, Raushel FM: **Evolution of function in (β/α)8-barrel enzymes.** *Curr Opin Chem Biol* 2003, **7**:252-264.

32. Höcker B, Claren J, Sterner R: **Mimicking enzyme evolution by generating new (betaalpha)8-barrels from (betaalpha)4-half-barrels**. *Proc Natl Acad Sci U S A* 2004, **10**:16448-16453.

33. Claren J, Malisi C, Höcker B, Sterner R: **Establishing wild-type levels of catalytic activity on natural and artificial (βα)8-barrel protein scaffolds**. *Proc Natl Acad Sci U S A* 2009, **106**:3704-3709.

34. Höcker B, Beismann-Driemeyer S, Hettwer S, Lustig A, Sterner R: **Dissection of a (βα)8-barrel enzyme into two folded halves**. *Nat Struct Biol* 2001, **8**:32-36.

35. Seitz T, Bocola M, Claren J, Sterner R: **Stabilisation of a (βα)8-barrel protein designed from identical half barrels**. *J Mol Biol* 2007, **372**:114-129.

36. Sharma P, Kaila P, Guptasarma P: **Creation of active TIM barrel enzymes through genetic fusion of half-barrel domain constructs derived from two distantly related glycosyl hydrolases**. *FEBS J* 2016, **283**:4340-4356.

37. Almeida VM, Frutuoso MA, Marana SR: **Search for independent (β/α)4 subdomains in a (β/α)8 barrel β-glucosidase**. *PLoS One* 2018, **13**:e0191282.

38. Wang JJ, Zhang T, Liu R, Song M, Wang JJ, Hong J, Chen Q, Liu H: **Recurring sequence-structure motifs in (βα)8-barrel proteins and experimental optimization of a chimeric protein designed based on such motifs**. *Biochim Biophys Acta - Proteins Proteomics* 2017, **1865**:165-175.

39. •• Lapidoth G, Khersonsky O, Lipsh R, Dym O, Albeck S, Rogotner S, Fleishman SJ: **Highly active enzymes by automated combinatorial backbone assembly and sequence design**. *Nat Commun* 2018, **9**:2780
A design approach is reported that uses protein fragments to create combinatorial backbones by applying evolutionary mechanisms during the assembly. This is applied to two homologous TIM-barrel families to build enzymes, which either restore or even increase the parental activities.

40. • ElGamacy M, Coles M, Lupas A: **Asymmetric protein design from conserved supersecondary structures**. *J Struct Biol* 2018, **204**:380-387
The authors explored the combination of heterologous structural motifs as a new potential mechanism in protein fold evolution. Replacement of an αα-hairpin motif from an unrelated family followed by interface optimization led to a stable, well-folded dRP-lyase protein verified by NMR.

41. Bharat TAM, Eisenbeis S, Zeth K, Höcker B: **A βα-barrel built by the combination of fragments from different folds**. *Proc Natl Acad Sci U S A* 2008, **105**:9942-9947.

42. Eisenbeis S, Proffitt W, Coles M, Truffault V, Shanmugaratnam S, Meiler J, Höcker B: **Potential of fragment recombination for rational design of proteins**. *J Am Chem Soc* 2012, **134**:4019-4022.

43. Shanmugaratnam S, Eisenbeis S, Höcker B: **A highly stable protein chimera built from fragments of different folds**. *Protein Eng Des Sel* 2012, **25**:699-703.

44. • Toledo-Patiño S, Chaubey M, Coles M, Höcker B: **Reconstructing the remote origins of a fold singleton from a flavodoxin-like ancestor**. *Biochemistry* 2019, **58**:4790-4793
A hypothetical evolutionary pathway is reconstructed supporting the emergence of the HemD-like fold from the flavodoxin-like fold. This work exemplifies how protein fold evolution can be reconstructed in the lab using modern day proteins.

45. Ferruz N, Noske J, Höcker B: **Protlego: a python package for the analysis and design of chimeric proteins**. *bioRxiv* 2020 http://dx.doi.org/10.1101/2020.10.04.325555.

46. Harms MJ, Thornton JW: **Evolutionary biochemistry: revealing the historical and physical causes of protein properties**. *Nat Rev Genet* 2013, **14**:559-571.

47. •• Halloran KT, Wang Y, Arora K, Chakravarthy S, Irving TC, Bilsel O, Brooks CL, Robert Matthews C: **Frustration and folding of a TIM barrel protein**. *Proc Natl Acad Sci U S A* 2019, **116**:16378-16383
This work combines computational simulations with experiments to monitor the folding pathway of indole-3-glycerol phosphate synthase. A common structurally-stable intermediate is formed early in folding and it can be concluded that the rate-limiting step in the folding pathway is the closing of the β-barrel.

48. Romero-Romero S, Becerril-Sesín LA, Costas M, Rodríguez-Romero A, Fernández-Velasco DA: **Structure and conformational stability of the triosephosphate isomerase from Zea mays. Comparison with the chemical unfolding pathways of other eukaryotic TIMs**. *Arch Biochem Biophys* 2018, **658**:66-76.

49. • Goldman AD, Beatty JT, Landweber LF: **The TIM barrel architecture facilitated the early evolution of protein-mediated metabolism**. *J Mol Evol* 2016, **82**:17-26
This bioinformatic work studies the role of the TIM-barrel fold in early evolution. Analysis of function, cofactor usage, and metabolic pathways suggested that TIM-barrel proteins participated in the transition from non-peptidic catalysis to protein-mediated metabolism.

50. • Richard JP: **Protein flexibility and stiffness enable efficient enzymatic catalysis**. *J Am Chem Soc* 2019, **141**:3320-3331
The author discusses how protein flexibility is an important factor in TIM-barrel enzymes to find a balance between substrate-binding energy and catalysis. It is suggested that selection of flexible loops provides a starting point for the evolution and divergence of new enzyme activities.

51. Quezada AG, Díaz-Salazar AJ, Cabrera N, Pérez-Montfort R, Piñeiro Á, Costas M: **Interplay between protein thermal flexibility and kinetic stability**. *Structure* 2017, **25**:167-179.

52. • Quezada AG, Cabrera N, Piñeiro Á, Díaz-Salazar AJ, Díaz-Mazariegos S, Romero-Romero S, Pérez-Montfort R, Costas M: **A strategy based on thermal flexibility to design triosephosphate isomerase proteins with increased or decreased kinetic stability**. *Biochem Biophys Res Commun* 2018, **503**:3017-3022
Two closely-related triosephosphate isomerases are used to explore the correlation between protein thermal flexibility and kinetic stability. Based on MD simulations and DSC experiments, the authors designed new functional TIM-barrel enzymes with fine-tuned kinetic stabilities.

53. Smith JM: **Natural selection and the concept of a protein space.** *Nature* 1970, **225**:563-564.

54. •• Chan YH, Venev SV, Zeldovich KB, Matthews CR: **Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints**. *Nat Commun* 2017, **8**:1-12
A mutational scanning approach is used to analyze the conservation of the fitness landscapes in three orthologous indole-3-glycerol phosphate synthases. It was found that their fitness landscapes are correlated, influenced by epistasis, and translocate sequence space due to the plasticity of the TIM-barrel fold.

55. Chan YH, Zeldovich KB, Matthews CR: **An allosteric pathway explains beneficial fitness in yeast for long-range mutations in an essential TIM barrel enzyme.** *Protein Sci* 2020, **29**:1911-1923.

56. Tzul FO, Vasilchuk D, Makhatadze GI: **Evidence for the principle of minimal frustration in the evolution of protein folding landscapes**. *Proc Natl Acad Sci U S A* 2017, **114**:E1627-E1632.

57. Candel AM, Romero-Romero ML, Gamiz-Arco G, Ibarra-Molero B, Sanchez-Ruiz JM: **Fast folding and slow unfolding of a resurrected Precambrian protein**. *Proc Natl Acad Sci U S A* 2017, **114**:E4122-E4123.

58. Gamiz-Arco G, Risso VA, Candel AM, Inglés-Prieto A, Romero-Romero ML, Gaucher EA, Gavira JA, Ibarra-Molero B, Sanchez-Ruiz JM: **Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding.** *Biochem J* 2019, **476**:3631-3647.

59. Kukic P, Pustovalova Y, Camilloni C, Gianni S, Korzhnev DM, Vendruscolo M: **Structural characterization of the early events in the nucleation-condensation mechanism in a protein folding process**. *J Am Chem Soc* 2017, **139**:6899-6910.

60. Hart KM, Harms MJ, Schmidt BH, Elya C, Thornton JW: **Thermodynamic system drift in protein evolution.** *PLoS Biol* 2014, **12** 1001994.

61. Lim, Shion A, Marqusee S: **The burst-phase folding intermediate of ribonuclease H changes conformation over evolutionary history**. *Biopolymers* 2018, **109**:e23086.

62. Lim SA, Hart KM, Harms MJ, Marqusee S: **Evolutionary trend toward kinetic stability in the folding trajectory of RNases H**. *Proc Natl Acad Sci U S A* 2016, **113**:13045-13050.

63. Lim SA, Bolin ER, Marqusee S: **Tracing a protein's folding pathway over evolutionary time using ancestral sequence reconstruction and hydrogen exchange**. *eLife* 2018, **7**:e38369.

64. Yang J, Naik N, Patel JS, Wylie CS, Gu W, Huang J, Marty Ytreberg F, Naik MT, Weinreich DM, Rubenstein BM: **Predicting the viability of beta-lactamase: how folding and binding free**

energies correlate with beta-lactamase fitness. *PLoS One* 2020, **15**:e0233509.

65. Faber MS, Wrenbeck EE, Azouz LR, Steiner PJ, Whitehead TA: **Impact of in vivo protein folding probability on local fitness landscapes.** *Mol Biol Evol* 2019, **36**:2764-2777.

66. Choi H-K, Min D, Kang H, Ju Shon M, Rah S-H, Chan Kim H, Jeong H, Choi H-J, Bowie JU, Yoon T-Y: **Watching helical membrane proteins fold reveals a common N-to-C-terminal folding pathway.** *Science* 2019, **366**:1150-1156.

67. Isogai Y, Imamura H, Nakae S, Sumi T, Takahashi KI, Nakagawa T, Tsuneshige A, Shirai T: **Tracing whale myoglobin evolution by resurrecting ancient proteins**. *Sci Rep* 2018, **8** 16883.

68. Korendovych IV, DeGrado WF: **De novo protein design, a retrospective.** *Q Rev Biophys* 2020, **53**:e3.

69. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W *et al.*: **Rosetta3: an object-oriented software suite for the simulation and design of macromolecules.** *Methods Enzymol* 2011, **487**:545-574.

70. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K *et al.*: **The rosetta all-atom energy function for macromolecular modeling and design**. *J Chem Theory Comput* 2017, **13**:3031-3048.

71. Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T, Kuhlman B: **Design of structurally distinct proteins using strategies inspired by evolution.** *Science* 2016, **352**:687-690.

72. Zhou J, Panaitiu AE, Grigoryan G: **A general-purpose protein design framework based on mining sequence-structure relationships in known protein structures**. *Proc Natl Acad Sci U S A* 2020, **117**:1059-1068.

73. Koga R, Koga N: **Consistency principle for protein design**. *Biophys Physicobiology* 2019, **16**:304-309.

74. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D: **Principles for designing ideal protein structures.** *Nature* 2012, **491**:222-227.

75. Goraj K, Renard A, Martial JA: **Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modelled on the α/β-barrel proteins.** *Protein Eng Des Sel* 1990, **3**:259-266.

76. Tanaka T, Hayashi M, Kimura H, Oobatake M, Nakamura H: **De novo design and creation of a stable artificial protein**. *Biophys Chem* 1994, **50**:47-61.

77. Tanaka T, Kuroda Y, Kimura H, Kidokoro SI, Nakamura H: **Cooperative deformation of a de novo designed protein.** *Protein Eng Des Sel* 1994, **7**:969-976.

78. Tanaka T, Kimura H, Hayashi M, Fujiyoshi Y, Fukuhara KI, Nakamura H: **Characteristics of a de novo designed protein.** *Protein Sci* 1994, **3**:419-427.

79. Houbrechts A, Moreau B, Abagyan R, Mainfroid V, Préaux G, Lamproye A, Poncin A, Goormaghtigh E, Ruysschaert JM, Martial JA *et al.*: **Second-generation octarellins: Two new de novo (β/α)8 polypeptides designed for investigating the influence of β-residue packing on the α/β-barrel structure stability**. *Protein Eng Des Sel* 1995, **8**:249-259.

80. Beauregard M, Goraj K, Goffin V, Heremans K, Goormaghtigh E, Ruysschaert JM, Martial JA: **Spectroscopic investigation of structure in octarellin (a de novo protein designed to adopt the α/β-barred packing)**. *Protein Eng Des Sel* 1991, **4**:745-749.

81. Offredi F, Dubail F, Kischel P, Sarinski K, Stern AS, Van de Weerdt C, Hoch JC, Prosperi C, François JM, Mayo SL *et al.*: **De novo backbone and sequence design of an idealized α/β-barrel protein: evidence of stable tertiary structure**. *J Mol Biol* 2003, **325**:163-174.

82. Figueroa M, Oliveira N, Lejeune A, Kaufmann KW, Dorr BM, Matagne A, Martial JA, Meiler J, Van de Weerdt C: **Octarellin VI: using rosetta to design a putative artificial (β/α)8 protein**. *PLoS One* 2013, **8**:e71858.

83. Figueroa M, Sleutel M, Vandevenne M, Parvizi G, Attout S, Jacquin O, Vandenameele J, Fischer AW, Damblon C, Goormaghtigh E *et al.*: **The unexpected structure of the designed protein Octarellin V.1 forms a challenge for protein structure prediction tools**. *J Struct Biol* 2016, **195**:19-30.

84. Nagarajan D, Deka G, Rao M: **Design of symmetric TIM barrel proteins from first principles.** *BMC Biochem* 2015, **16**:18.

85. • Huang PS, Feldmeier K, Parmeggiani F, Fernandez Velasco DA, Hocker B, Baker D: **De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy**. *Nat Chem Biol* 2016, **12**:29-34

This work reports the first successful *de novo* designed TIM barrel. The design approach comprises determination of geometric restrictions, backbone generation and iterative sequence design. Experimental characterization showed well-folded proteins and the intended topology for the construct sTIM11.

86. •• Romero-Romero S, Costas M, Silva D-A, Kordes S, Rojas-Ortega E, Tapia C, Guerra Y, Shanmugaratnam S, Rodríguez-Romero A, Baker D *et al.*: **Epistasis on the stability landscape of de novo TIM barrels explored by a modular design approach.** *bioRxiv* 2020 http://dx.doi.org/10.1101/2020.09.29.319103

A modular design approach was used to create a family of stabilized sTIM11 variants by improving hydrophobic packing. Detailed analysis showed that unexplored regions of the stability landscape are accessed. This landscape is shaped by epistatic effects arising from improved hydrophobic clusters.

87. Caldwell SJ, Haydon IC, Piperidou N, Huang P-S, Bick MJ, Sjöström HS, Hilvert D, Baker D, Zeymer C: **Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion.** *Proc Natl Acad Sci U S A* 2020, **117**:30362-30369.

88. Wiese G, Shanmugaratnam S, Höcker B: **Extension of a de novo TIM barrel with a rationally designed secondary structure element.** *bioRxiv* 2020 http://dx.doi.org/10.1101/2020.10.16.342774.

89. Watters AL, Deka P, Corrent C, Callender D, Varani G, Sosnick T, Baker D: **The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection**. *Cell* 2007, **128**:613-624.

90. •• Basak S, Paul Nobrega R, Tavella D, Deveau LM, Koga N, Tatsumi-Koga R, Baker D, Massi F, Robert Matthews C: **Networks of electrostatic and hydrophobic interactions modulate the complex folding free energy surface of a designed βα protein.** *Proc Natl Acad Sci U S A* 2019, **116**:6806-6811

This work analyses the complex folding pathway of the *de novo* protein Di-III_14. Electrostatic and hydrophobic networks are identified as possible modulators and their contribution specified by mutational analysis. These findings have implications for future protein design strategies.

9. List of Publications

The following publications are the underlying scientific studies composing this thesis:

I. **Retracing the Evolution of a Modern Periplasmic Binding Protein**
Florian Michel*, Sergio Romero-Romero*, Birte Höcker

Protein Science, 2023, 32(11)
https://doi.org/10.1002/pro.4793

II. **Structures of Permuted Halves of a Modern Ribose Binding Protein**
Florian Michel, Sooruban Shanmugaratnam, Sergio Romero-Romero, Birte Höcker

Acta Crystallographica Section D, 2023, D79, 40-49
https://doi.org/10.1107/S205979832201186X

III. **Fuzzle 2.0: Ligand Binding in Natural Protein Building Blocks**
Noelia Ferruz, **Florian Michel**, Francisco Lobos, Steffen Schmidt, Birte Höcker

Frontiers in Molecular Biosciences. 2021, 18;8:715972
https://doi.org/10.3389/fmolb.2021.715972

IV. **Isolation of Subdomain-sized Elements in a Modern Periplasmic Binding Protein**
Florian Michel, Timo Kossendey, Birte Höcker

Manuscript

V. **Evolution, Folding and Design of TIM Barrels and Related Proteins**
Sergio Romero-Romero*, Sina Kordes*, **Florian Michel***, Birte Höcker

Current Opinion in Structural Biology, 2021, 68, pp. 94-104

https://doi.org/10.1016/j.sbi.2020.12.007

* equal contribution

10. Outlook and perspective

This work tries to establish a more complete view on the evolution of protein folds. In this journey “back in time”, we start with the modern ribose binding protein of *T.maritima*, a protein adapted to bind periplasmatic ribose with a high affinity. Following up on the idea that this specific protein fold stems from a duplication, the modern protein was then disassembled into its two constituent parts. It was possible to isolate both lobes of the protein and investigate their structure, showing that they closely resemble the proposed flavodoxin-like topology thought to be its ancestor. Furthermore, it could be proven that the disassembled protein retains almost native-like function when the two individual lobes are in presence of each other.

Using this system to go even further back, we identified a fragment of roughly 90 residues within RBP. Since such fragments are believed to be remnants of ancient elements present at the origin of protein evolution, an attempt to isolate this part of the protein was made. The successful purification and characterization of this fragment lends credibility to the idea that these fragments pose important structural elements even in modern protein folds. Their remarkable interconnectivity and distribution in the modern protein fold space helps us understand universal rules of these elements. Possible application of this knowledge in evolutionary-informed protein design could be an easy and suitable way to circumvent extensive and computationally expensive protein design *de-novo* in the future.

Combining this existing dataset of fragments found in Fuzzle with the additional feature of analyzing ligand binding within fragments can also help in constructing functional chimeras using this mix-and-match approach to design proteins with new abilities.

While the evolutionary study of the PBP-like fold is by no means meant to be comprehensive, it adds additional credibility to the theory of duplication. The studies on the stability of the fragments might inspire others to seek similar structural elements in their proteins, and possibly broaden how we view the role of these subdomain fragments in the course of evolution.

11. Literature

(1) Chothia, C.; Gough, J. Genomic and Structural Aspects of Protein Evolution. *Biochemical Journal*. 2009, pp 15–28. https://doi.org/10.1042/BJ20090122.

(2) ANFINSEN, C. B.; HABER, E.; SELA, M.; WHITE, F. H. The Kinetics of Formation of Native Ribonuclease during Oxidation of the Reduced Polypeptide Chain. *Proc. Natl. Acad. Sci. U. S. A.* **1961**, *47* (9), 1309–1314. https://doi.org/10.1073/PNAS.47.9.1309/ASSET/D9AA8804-3797-4BF3-8045-576A3ACD0DE7/ASSETS/PNAS.47.9.1309.FP.PNG.

(3) Levinthal, C. How to Fold Graciously. In *Mössbaun Spectroscopy in Biological Systems Proceedings*; 1969. https://doi.org/http://www.cc.gatech.edu/~turk/bio_sim/articles/proteins_levinthal_1969.pdf.

(4) *CSHL Archives Repository* | *On Protein Synthesis*. http://libgallery.cshl.edu/items/show/52220 (accessed 2022-11-04).

(5) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29* (31), 7133–7155. https://doi.org/10.1021/BI00483A001/ASSET/BI00483A001.FP.PNG_V03.

(6) Kubelka, J.; Hofrichter, J.; Eaton, W. A. The Protein Folding 'Speed Limit.' *Curr. Opin. Struct. Biol.* **2004**, *14* (1), 76–88. https://doi.org/10.1016/J.SBI.2004.01.013.

(7) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science (80-.).* **2011**, *334* (6055), 517–520. https://doi.org/10.1126/SCIENCE.1208351/SUPPL_FILE/LINDORFF-LARSEN_SOM-REVISION1.PDF.

(8) Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. Is Burst Hydrophobic Collapse Necessary for Protein Folding? *Biochemistry* **1995**, *34* (9), 3066–3076. https://doi.org/10.1021/BI00009A038/ASSET/BI00009A038.FP.PNG_V03.

(9) Christensen, H.; Pain, R. H. Molten Globule Intermediates and Protein Folding. *Eur. Biophys. J. 1991 195* **1991**, *19* (5), 221–229. https://doi.org/10.1007/BF00183530.

(10) Kuwajima, K. The Molten Globule State as a Clue for Understanding the Folding and Cooperativity of Globular-Protein Structure. *Proteins Struct. Funct. Bioinforma.* **1989**, *6* (2), 87–103. https://doi.org/10.1002/PROT.340060202.

(11) Sadqi, M.; Lapidus, L. J.; Muñoz, V. How Fast Is Protein Hydrophobic Collapse? *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (21), 12117–12122. https://doi.org/10.1073/PNAS.2033863100/SUPPL_FILE/3863FIG7.PDF.

(12) Radford, S. E.; Dobson, C. M.; Evans, P. A. The Folding of Hen Lysozyme Involves Partially Structured Intermediates and Multiple Pathways. *Nat. 1992 3586384* **1992**, *358* (6384), 302–307. https://doi.org/10.1038/358302a0.

(13) Matagne, A.; Dobson, C. M. The Folding Process of Hen Lysozyme: A Perspective from the 'New View.' *Cell. Mol. Life Sci. C. 1998 544* **2014**, *54* (4), 363–371. https://doi.org/10.1007/S000180050165.

(14) Nagano, N.; Orengo, C. A.; Thornton, J. M. One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on Their Sequences, Structures and Functions. *J. Mol. Biol.* **2002**, *321* (5), 741–765. https://doi.org/10.1016/S0022-2836(02)00649-6.

(15) Buckle, A. M.; Schreiber, G.; Fersht, A. R. Protein-Protein Recognition: Crystal Structural Analysis of a Barnase-Barstar Complex at 2.0-Å Resolution. *Biochemistry* **1994**, *33* (30), 8878–8889. https://doi.org/10.1021/BI00196A004/ASSET/BI00196A004.FP.PNG_V03.

(16) Golbik, R.; Fischer, G.; Fersht, A. R. Folding of Barstar C40A/C82A/P27A and Catalysis of the Peptidyl-Prolyl Cis/Trans Isomerization by Human Cytosolic Cyclophilin (Cyp18). *Protein Sci.* **1999**, *8* (7), 1505. https://doi.org/10.1110/PS.8.7.1505.

(17) Agashe, V. R.; Shastry, M. C. R.; Udgaonkar, J. B. Initial Hydrophobic Collapse in the Folding of Barstar. *Nat. 1995 3776551* **1995**, *377* (6551), 754–757. https://doi.org/10.1038/377754a0.

(18) Schmid, F. X.; Baldwin, R. L. Detection of an Early Intermediate in the Folding of Ribonuclease A by Protection of Amide Protons against Exchange. *J. Mol. Biol.* **1979**, *135* (1), 199–215. https://doi.org/10.1016/0022-2836(79)90347-4.

(19) Gilmanshin, R.; Dyer, R. B.; Callender, R. H. Structural Heterogeneity of the Various Forms of Apomyoglobin; Implications for Protein Folding. *Protein Sci.* **1997**, *6* (10), 2134–2142. https://doi.org/10.1002/PRO.5560061008.

(20) Ptitsyn, O. B.; Rashin, A. A. A Model of Myoglobin Self-Organization. *Biophys. Chem.* **1975**, *3* (1), 1–20. https://doi.org/10.1016/0301-4622(75)80033-0.

(21) Steegborn, C.; Schneider-Hassloff, H.; Zeeb, M.; Balbach, J. Cooperativity of a Protein Folding Reaction Probed at Multiple Chain Positions by Real-Time 2D NMR Spectroscopy†. *Biochemistry* **2000**, *39* (27), 7910–7919. https://doi.org/10.1021/BI000270U.

(22) Karplus, M.; Weaver, D. L. Protein Folding Dynamics: The Diffusion-Collision Model and Experimental Data. *Protein Sci.* **1994**, *3* (4), 650–668. https://doi.org/10.1002/PRO.5560030413.

(23) Kim, P. S.; Baldwin, R. L. INTERMEDIATES IN THE FOLDING REACTIONS OF SMALL PROTEINS. *https://doi.org/10.1146/annurev.bi.59.070190.003215* **2003**, *59* (1), 631–660. https://doi.org/10.1146/ANNUREV.BI.59.070190.003215.

(24) Wetlaufer, D. B. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proc. Natl. Acad. Sci.* **1973**, *70* (3), 697–701. https://doi.org/10.1073/PNAS.70.3.697.

(25) Ptitsyn, O. B. Molten Globule and Protein Folding. *Adv. Protein Chem.* **1995**, *47*, 83–229. https://doi.org/10.1016/S0065-3233(08)60546-X.

(26) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science (80-.).* **1991**, *254* (5038), 1598–1603. https://doi.org/10.1126/SCIENCE.1749933.

(27) Honig, B.; Yang, A. S. Free Energy Balance in Protein Folding. *Adv. Protein Chem.* **1995**, *46* (C), 27–58. https://doi.org/10.1016/S0065-3233(08)60331-9.

(28) Ferenczy, G. G.; Kellermayer, M. Contribution of Hydrophobic Interactions to Protein Mechanical Stability. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 1946–1956. https://doi.org/10.1016/j.csbj.2022.04.025.

(29) Taverna, D. M.; Goldstein, R. A. Why Are Proteins Marginally Stable? *Proteins Struct. Funct. Genet.* **2002**, *46* (1), 105–109. https://doi.org/10.1002/prot.10016.

(30) Goldstein, R. A. The Evolution and Evolutionary Consequences of Marginal Thermostability in Proteins. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (5), 1396–1407. https://doi.org/10.1002/prot.22964.

(31) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103* (2), 227–249. https://doi.org/10.1016/0022-2836(76)90311-9.

(32) McLendon, G.; Radany, E. Is Protein Turnover Thermodynamically Controlled? *J. Biol. Chem.* **1978**, *253* (18), 6335–6337. https://doi.org/10.1016/s0021-9258(19)46935-4.

(33) Parsell, D. A.; Sauer, R. T. The Structural Stability of a Protein Is an Important Determinant of Its Proteolytic Susceptibility in Escherichia Coli. *J. Biol. Chem.* **1989**, *264* (13), 7590–7595. https://doi.org/10.1016/s0021-9258(18)83275-6.

(34) Lindberg, M.; Tångrot, J.; Oliveberg, M. Complete Change of the Protein Folding Transition State upon Circular Permutation. *Nat. Struct. Biol.* **2002**, *9* (11), 818–822. https://doi.org/10.1038/nsb847.

(35) Scalley-Kim, M.; Baker, D. Characterization of the Folding Energy Landscapes of Computer Generated Proteins Suggests High Folding Free Energy Barriers and Cooperativity May Be Consequences of Natural Selection. *J. Mol. Biol.* **2004**, *338* (3), 573–583. https://doi.org/10.1016/j.jmb.2004.02.055.

(36) Huang, P. S.; Boyken, S. E.; Baker, D. The Coming of Age of de Novo Protein Design. *Nature*. Nature Publishing Group September 14, 2016, pp 320–327. https://doi.org/10.1038/nature19946.

(37) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science (80-.).* **2003**, *302* (5649), 1364–1368. https://doi.org/10.1126/SCIENCE.1089427/SUPPL_FILE/1089427S.PDF.

(38) Romero-Romero, S.; Costas, M.; Silva Manzano, D. A.; Kordes, S.; Rojas-Ortega, E.; Tapia, C.; Guerra, Y.; Shanmugaratnam, S.; Rodríguez-Romero, A.; Baker, D.; Höcker, B.; Fernández-Velasco, D. A. The Stability Landscape of de Novo TIM Barrels Explored by a Modular Design Approach. *J. Mol. Biol.* **2021**, *433* (18), 167153. https://doi.org/10.1016/J.JMB.2021.167153.

(39) Goldenzweig, A.; Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *https://doi.org/10.1146/annurev-biochem-062917-012102* **2018**, *87*, 105–129. https://doi.org/10.1146/ANNUREV-BIOCHEM-062917-012102.

(40) Benner, S. A.; Gaucher, E. A. Evolution, Language and Analogy in Functional Genomics. *Trends Genet.* **2001**, *17* (7), 414–418. https://doi.org/10.1016/S0168-9525(01)02320-4.

(41) Sereno, M. I. Four Analogies between Biological and Cultural/Linguistic Evolution. *J. Theor. Biol.* **1991**, *151* (4), 467–507. https://doi.org/10.1016/S0022-5193(05)80366-2.

(42) Alva, V.; Söding, J.; Lupas, A. N. A Vocabulary of Ancient Peptides at the Origin of Folded Proteins. *Elife* **2015**, *4* (DECEMBER2015), 1–19. https://doi.org/10.7554/eLife.09410.001.

(43) Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nat. 1970 2255232* **1970**, *225* (5232), 563–564. https://doi.org/10.1038/225563a0.

(44) Brandon Ogbunugafor, C. A Reflection on 50 Years of John Maynard Smith's "Protein Space." *Genetics* **2020**, *214* (4), 749–754. https://doi.org/10.1534/GENETICS.119.302764.

(45) Caetano-Anollés, G.; Kim, K. M.; Caetano-Anollés, D. The Phylogenomic Roots of Modern Biochemistry: Origins of Proteins, Cofactors and Protein Biosynthesis. *J. Mol. Evol.* **2012**, *74* (1–2), 1–34. https://doi.org/10.1007/s00239-011-9480-1.

(46) Kinch, L. N.; Grishin, N. V. Evolution of Protein Structures and Functions. *Curr. Opin. Struct. Biol.* **2002**, *12* (3), 400–408. https://doi.org/10.1016/S0959-440X(02)00338-X.

(47) Höcker, B.; Winther, J. R. Editorial Overview: A Perspective on Protein Evolution. *Curr. Opin. Struct. Biol.* **2018**, *48*, viii–ix. https://doi.org/10.1016/J.SBI.2018.01.013.

(48) Höcker, B. Design of Proteins from Smaller Fragments — Learning from Evolution. *Curr. Opin. Struct. Biol.* **2014**, *27* (1), 56–62. https://doi.org/10.1016/J.SBI.2014.04.007.

(49) Eisenbeis, S.; Höcker, B. Evolutionary Mechanism as a Template for Protein Engineering. *J. Pept. Sci.* **2010**, *16* (10), 538–544. https://doi.org/10.1002/PSC.1233.

(50) Ferruz, N.; Höcker, B. Controllable Protein Design with Language Models. *Nat. Mach. Intell. 2022 46* **2022**, *4* (6), 521–532. https://doi.org/10.1038/s42256-022-00499-z.

(51) Levinthal, C. Are There Pathways for Protein Folding? *J. Chim. Phys.* **1968**, *65*, 44–45. https://doi.org/10.1051/JCP/1968650044.

(52) Dill, K. A.; Alonso, D. O. V.; Hutchinson, K. Thermal Stabilities of Globular Proteins. *Biochemistry* **1989**, *28* (13), 5439–5449. https://doi.org/10.1021/BI00439A019/ASSET/BI00439A019.FP.PNG_V03.

(53) Dill, K. A.; Chan, H. S. From Levinthal to Pathways to Funnels. **1997**.

(54) Ptitsyn, O. B. Protein Folding: Hypotheses and Experiments. *J. Protein Chem. 1987 64* **1987**, *6* (4), 273–293. https://doi.org/10.1007/BF00248050.

(55) Hingorani, K. S.; Gierasch, L. M. Comparing Protein Folding in Vitro and in Vivo: Foldability Meets the Fitness Challenge. *Curr. Opin. Struct. Biol.* **2014**, *24* (1), 81–90. https://doi.org/10.1016/J.SBI.2013.11.007.

(56) Gruebele, M.; Dave, K.; Sukenik, S. Globular Protein Folding In Vitro and In Vivo. *https://doi.org/10.1146/annurev-biophys-062215-011236* **2016**, *45*, 233–251. https://doi.org/10.1146/ANNUREV-BIOPHYS-062215-011236.

(57) Hartl, F. U.; Hayer-Hartl, M. Converging Concepts of Protein Folding in Vitro and in Vivo. *Nat. Struct. Mol. Biol. 2009 166* **2009**, *16* (6), 574–581. https://doi.org/10.1038/nsmb.1591.

(58) Lu, H.; Diaz, D. J.; Czarnecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; Lynd, N. A.; Ellington, A. D.; Alper, H. S. Machine Learning-Aided Engineering of Hydrolases for PET Depolymerization. *Nat. 2022 6047907* **2022**, *604* (7907), 662–667. https://doi.org/10.1038/s41586-022-04599-z.

(59) Pearce, R.; Zhang, Y. Deep Learning Techniques Have Significantly Impacted Protein Structure Prediction and Protein Design. *Curr. Opin. Struct. Biol.* **2021**, *68*, 194–207. https://doi.org/10.1016/J.SBI.2021.01.007.

(60) Pan, X.; Kortemme, T. Recent Advances in de Novo Protein Design: Principles, Methods, and Applications. *J. Biol. Chem.* **2021**, *296*, 100558. https://doi.org/10.1016/J.JBC.2021.100558.

(61) Woolfson, D. N. A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. *J. Mol. Biol.* **2021**, *433* (20), 167160. https://doi.org/10.1016/J.JMB.2021.167160.

(62) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Santos Costa, A. Dos; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, [2] Alexander. Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model. *bioRxiv* **2022**, 2022.07.20.500902. https://doi.org/10.1101/2022.07.20.500902.

(63) McLachlan, A. D. Repeating Sequences and Gene Duplication in Proteins. *J. Mol. Biol.* **1972**, *64* (2), 417–437. https://doi.org/10.1016/0022-2836(72)90508-6.

(64) Lupas, A. N.; Ponting, C. P.; Russell, R. B. 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? *J. Struct. Biol.* **2001**, *134* (2–3), 191–203. https://doi.org/10.1006/jsbi.2001.4393.

(65) Söding, J.; Lupas, A. N. More than the Sum of Their Parts: On the Evolution of Proteins from Peptides. *BioEssays* **2003**, *25* (9), 837–846. https://doi.org/10.1002/bies.10321.

(66) Alva, V.; Ammelburg, M.; Söding, J.; Lupas, A. N. On the Origin of the Histone Fold. *BMC Struct. Biol.* **2007**, *7*, 1–10. https://doi.org/10.1186/1472-6807-7-17.

(67) Alva, V.; Lupas, A. N. From Ancestral Peptides to Designed Proteins. *Curr. Opin. Struct. Biol.* **2018**, *48* (December), 103–109. https://doi.org/10.1016/j.sbi.2017.11.006.

(68) Andreeva, A.; Howorth, D.; Chothia, C.; Kulesha, E.; Murzin, A. G. SCOP2 Prototype: A New Approach to Protein Structure Mining. *Nucleic Acids Res.* **2014**, *42* (D1), D310–D314. https://doi.org/10.1093/NAR/GKT1242.

(69) Andreeva, A.; Kulesha, E.; Gough, J.; Murzin, A. G. The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures. *Nucleic Acids Res.* **2020**, *48* (D1), D376–D382. https://doi.org/10.1093/NAR/GKZ1064.

(70) Sillitoe, I.; Dawson, N.; Lewis, T. E.; Das, S.; Lees, J. G.; Ashford, P.; Tolulope, A.; Scholes, H. M.; Senatorov, I.; Bujan, A.; Ceballos Rodriguez-Conde, F.; Dowling, B.; Thornton, J.; Orengo, C. A. CATH: Expanding the Horizons of Structure-Based Functional Annotations for Genome Sequences. *Nucleic Acids Res.* **2019**, *47* (D1), D280–D284. https://doi.org/10.1093/NAR/GKY1097.

(71) Cheng, H.; Schaeffer, R. D.; Liao, Y.; Kinch, L. N.; Pei, J.; Shi, S.; Kim, B.-H.; Grishin, N. V. ECOD: An Evolutionary Classification of Protein Domains. *PLoS Comput. Biol.* **2014**, *10* (12), e1003926. https://doi.org/10.1371/journal.pcbi.1003926.

(72) Dustin Schaeffer, R.; Liao, Y.; Cheng, H.; Grishin, N. V. ECOD: New Developments in the Evolutionary Classification of Domains. *Nucleic Acids Res.* **2017**, *45* (D1), D296–D302. https://doi.org/10.1093/nar/gkw1137.

(73) Chandonia, J. M.; Guan, L.; Lin, S.; Yu, C.; Fox, N. K.; Brenner, S. E. SCOPe: Improvements to the Structural Classification of Proteins – Extended Database to Facilitate Variant Interpretation and Machine Learning. *Nucleic Acids Res.* **2022**, *50* (D1), D553–D559. https://doi.org/10.1093/NAR/GKAB1054.

(74) Hu, X.; Feng, C.; Ling, T.; Chen, M. Deep Learning Frameworks for Protein–Protein Interaction Prediction. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 3223–3233. https://doi.org/10.1016/J.CSBJ.2022.06.025.

(75) Prabantu, V. M.; Gadiyaram, V.; Vishveshwara, S.; Srinivasan, N. Understanding Structural Variability in Proteins Using Protein Structural Networks. *Curr. Res. Struct. Biol.* **2022**, *4*, 134–145. https://doi.org/10.1016/J.CRSTBI.2022.04.002.

(76) Han, K.; Liu, Y.; Xu, J.; Song, J.; Yu, D. J. Performing Protein Fold Recognition by Exploiting a Stack Convolutional Neural Network with the Attention Mechanism. *Anal. Biochem.* **2022**, *651*, 114695. https://doi.org/10.1016/J.AB.2022.114695.

(77) Zimmermann, L.; Stephens, A.; Nam, S. Z.; Rau, D.; Kübler, J.; Lozajic, M.; Gabler, F.; Söding, J.; Lupas, A. N.; Alva, V. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at Its Core. *J. Mol. Biol.* **2018**, *430* (15), 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.

(78) Nepomnyachiy, S.; Ben-Tal, N.; Kolodny, R. Complex Evolutionary Footprints Revealed in an Analysis of Reused Protein Segments of Diverse Lengths. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (44), 11703–11708. https://doi.org/10.1073/pnas.1707642114.

(79) Söding, J. Protein Homology Detection by HMM–HMM Comparison. *Bioinformatics* **2005**, *21* (7), 951–960. https://doi.org/10.1093/BIOINFORMATICS/BTI125.

(80) Ferruz, N.; Lobos, F.; Lemm, D.; Toledo-Patino, S.; Farías-Rico, J. A.; Schmidt, S.; Höcker, B. Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design. *J. Mol. Biol.* **2020**, *432* (13), 3898–3914. https://doi.org/10.1016/J.JMB.2020.04.013.

(81) Zheng, Z.; Goncearenco, A.; Berezovsky, I. N. Nucleotide Binding Database NBDB – a Collection of Sequence Motifs with Specific Protein-Ligand Interactions. *Nucleic Acids Res.* **2016**, *44* (D1), D301–D307. https://doi.org/10.1093/NAR/GKV1124.

(82) Nepomnyachiy, S.; Ben-Tal, N.; Kolodny, R. Complex Evolutionary Footprints Revealed in an Analysis of Reused Protein Segments of Diverse Lengths. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (44), 11703–11708. https://doi.org/10.1073/PNAS.1707642114/SUPPL_FILE/PNAS.1707642114.SAPP.PDF.

(83) Kolodny, R. Searching Protein Space for Ancient Sub-Domain Segments. *Curr. Opin. Struct. Biol.* **2021**, *68*, 105–112. https://doi.org/10.1016/J.SBI.2020.11.006.

(84) Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33* (7), 2302–2309. https://doi.org/10.1093/NAR/GKI524.

(85) Ohta, T. Mechanisms of Molecular Evolution. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* **2000**, *355* (1403), 1623–1626. https://doi.org/10.1098/RSTB.2000.0724.

(86) Sikosek, T.; Chan, H. S. Biophysics of Protein Evolution and Evolutionary Protein Biophysics. *J. R. Soc. Interface* **2014**, *11* (100). https://doi.org/10.1098/RSIF.2014.0419.

(87) Söding, J.; Lupas, A. N. More than the Sum of Their Parts: On the Evolution of Proteins from Peptides. *BioEssays* **2003**, *25* (9), 837–846. https://doi.org/10.1002/BIES.10321.

(88) Chen, J.; Guo, M.; Wang, X.; Liu, B. A Comprehensive Review and Comparison of Different Computational Methods for Protein Remote Homology Detection. *Brief. Bioinform.* **2018**, *19* (2), 231–244. https://doi.org/10.1093/BIB/BBW108.

(89) Farías-Rico, J. A.; Schmidt, S.; Höcker, B. Evolutionary Relationship of Two Ancient Protein Superfolds. *Nat. Chem. Biol.* **2014**, *10* (9), 710–715. https://doi.org/10.1038/nchembio.1579.

(90) Toledo-Patiño, S.; Chaubey, M.; Coles, M.; Höcker, B. Reconstructing the Remote Origins of a Fold Singleton from a Flavodoxin-Like Ancestor. *Biochemistry* **2019**, *58* (48), 4790–4793. https://doi.org/10.1021/ACS.BIOCHEM.9B00900/ASSET/IMAGES/LARGE/BI9B00900_0002.JPEG.

(91) Gunasekaran, K.; Eyles, S. J.; Hagler, A. T.; Gierasch, L. M. Keeping It in the Family: Folding Studies of Related Proteins. *Curr. Opin. Struct. Biol.* **2001**, *11* (1), 83–93. https://doi.org/10.1016/S0959-440X(00)00173-1.

(92) Alva, V.; Remmert, M.; Biegert, A.; Lupas, A. N.; Söding, J. A Galaxy of Folds. *Protein Sci.* **2010**, *19* (1), 124–130. https://doi.org/10.1002/pro.297.

(93) Matilla, M. A.; Ortega, Á.; Krell, T. The Role of Solute Binding Proteins in Signal Transduction. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1786–1805. https://doi.org/10.1016/j.csbj.2021.03.029.

(94) Dwyer, M. A.; Hellinga, H. W. Periplasmic Binding Proteins: A Versatile Superfamily for Protein Engineering. *Curr. Opin. Struct. Biol.* **2004**, *14* (4), 495–504. https://doi.org/10.1016/J.SBI.2004.07.004.

(95) Felder, C. B.; Graul, R. C.; Lee, A. Y.; Merkle, H. P.; Sadee, W. The Venus Flytrap of Periplasmic Binding Proteins: An Ancient Protein Module Present in Multiple Drug Receptors. *AAPS PharmSci 1999 12* **1999**, *1* (2), 7–26. https://doi.org/10.1208/PS010202.

(96) Broehan, G.; Kroeger, T.; Lorenzen, M.; Merzendorfer, H. Functional Analysis of the ATP-Binding Cassette (ABC) Transporter Gene Family of Tribolium Castaneum. *BMC Genomics* **2013**, *14* (1), 1–19. https://doi.org/10.1186/1471-2164-14-6/FIGURES/5.

(97) Xiong, J.; Feng, L.; Yuan, D.; Fu, C.; Miao, W. Genome-Wide Identification and Evolution of ATP-Binding Cassette Transporters in the Ciliate Tetrahymena Thermophila: A Case of Functional Divergence in a Multigene Family. *BMC Evol. Biol.* **2010**, *10* (1), 1–18. https://doi.org/10.1186/1471-2148-10-330/TABLES/4.

(98) Davidson, A. L.; Chen, J. ATP-Binding Cassette Transporters in Bacteria. *https://doi.org/10.1146/annurev.biochem.73.011303.073626* **2004**, *73*, 241–268. https://doi.org/10.1146/ANNUREV.BIOCHEM.73.011303.073626.

(99) Zeng, Y.; Charkowski, A. O. The Role of ATP-Binding Cassette Transporters in Bacterial Phytopathogenesis. *Phytopathology* **2021**, *111* (4), 600–610. https://doi.org/10.1094/PHYTO-06-20-0212-RVW/ASSET/IMAGES/LARGE/PHYTO-06-20-0212-RVWF7.JPEG.

(100) Srikant, S. Evolutionary History of ATP-Binding Cassette Proteins. *FEBS Lett.* **2020**, *594* (23), 3882–3897. https://doi.org/10.1002/1873-3468.13985.

(101) Armstrong, N.; Gouaux, E. Mechanisms for Activation and Antagonism of an AMPA-Sensitive Glutamate Receptor: Crystal Structures of the GluR2 Ligand Binding Core. *Neuron* **2000**, *28* (1), 165–181. https://doi.org/10.1016/S0896-6273(00)00094-5.

(102) Kröger, P.; Shanmugaratnam, S.; Ferruz, N.; Schweimer, K.; Höcker, B. A Comprehensive Binding Study Illustrates Ligand Recognition in the Periplasmic Binding Protein PotF. *Structure* **2021**, *29* (5), 433-443.e4. https://doi.org/10.1016/j.str.2020.12.005.

(103) Cuneo, M. J.; Beese, L. S.; Hellinga, H. W. Ligand-Induced Conformational Changes in a Thermophilic Ribose-Binding Protein. *BMC Struct. Biol.* **2008**, *8* (1), 50. https://doi.org/10.1186/1472-6807-8-50.

(104) Marvin, J. S.; Corcoran, E. E.; Hattangadi, N. A.; Zhang, J. V.; Gere, S. A.; Hellinga, H. W. The Rational Design of Allosteric Interactions in a Monomeric Protein and Its Applications to the Construction of Biosensors. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (9), 4366–4371. https://doi.org/10.1073/PNAS.94.9.4366/ASSET/2290ED2A-910C-481E-A1DB-7C5E86AAB3CA/ASSETS/GRAPHIC/PQ0970487004.JPEG.

(105) Salins, L. L. E.; Ware, R. A.; Ensor, C. M.; Daunert, S. A Novel Reagentless Sensing System for Measuring Glucose Based on the Galactose/Glucose-Binding Protein. *Anal. Biochem.* **2001**, *294* (1), 19–26. https://doi.org/10.1006/ABIO.2001.5131.

(106) Lorimier, R. M. De; Smith, J. J.; Dwyer, M. A.; Looger, L. L.; Sali, K. M.; Paavola, C. D.; Rizk, S. S.; Sadigov, S.; Conrad, D. W.; Loew, L.; Hellinga, H. W. Construction of a Fluorescent Biosensor Family. *Protein Sci.* **2002**, *11* (11), 2655–2675. https://doi.org/10.1110/PS.021860.

(107) Salins, L. L. E.; Goldsmith, E. S.; Ensor, C. M.; Daunert, S. A Fluorescence-Based Sensing System for the Environmental Monitoring of Nickel Using the Nickel Binding Protein from Escherichia Coli. *Anal. Bioanal. Chem. 2001 3721* **2001**, *372* (1), 174–180. https://doi.org/10.1007/S00216-001-1169-7.

(108) Zemerov, S. D.; Roose, B. W.; Farenhem, K. L.; Zhao, Z.; Stringer, M. A.; Goldman, A. R.; Speicher, D. W.; Dmochowski, I. J. 129Xe NMR-Protein Sensor Reveals Cellular Ribose Concentration. *Anal. Chem.* **2020**, *92* (19), 12817–12824. https://doi.org/10.1021/ACS.ANALCHEM.0C00967/SUPPL_FILE/AC0C00967_SI_001.PDF.

(109) Chandrasekaran, N. I.; Matheswaran, M. Unique Nonenzymatic Glucose Sensor Using a Hollow-Shelled Triple Oxide Mn−Cu−Al Nanocomposite. **2020**. https://doi.org/10.1021/acsomega.0c00417.

(110) Ko, W.; Kumar, R.; Kim, S.; Lee, H. S. Construction of Bacterial Cells with an Active Transport System for Unnatural Amino Acids. *ACS Synth. Biol.* **2019**, *8* (5), 1195–1203. https://doi.org/10.1021/ACSSYNBIO.9B00076/SUPPL_FILE/SB9B00076_SI_001.PDF.

(111) Zhu, J.; Pei, D. A LuxP-Based Fluorescent Sensor for Bacterial Autoinducer II. *ACS Chem. Biol.* **2008**, *3* (2), 110–119. https://doi.org/10.1021/CB7002048/SUPPL_FILE/CB7002048-FILE007.PDF.

(112) Kröger, P.; Shanmugaratnam, S.; Scheib, U.; Höcker, B. Fine-Tuning Spermidine Binding Modes in the Putrescine Binding Protein PotF. *J. Biol. Chem.* **2021**, *297* (6). https://doi.org/10.1016/j.jbc.2021.101419.

(113) Banda-Vázquez, J.; Shanmugaratnam, S.; Rodríguez-Sotres, R.; Torres-Larios, A.; Höcker, B.; Sosa-Peinado, A. Redesign of LAOBP to Bind Novel L-Amino Acid Ligands. *Protein Sci.* **2018**, *27* (5), 957–968. https://doi.org/10.1002/PRO.3403.

(114) Scheepers, G. H.; Lycklama a Nijeholt, J. A.; Poolman, B. An Updated Structural Classification of Substrate-Binding Proteins. *FEBS Lett.* **2016**, *590* (23), 4393–4401. https://doi.org/10.1002/1873-3468.12445.

(115) Fukami-Kobayashi, K.; Tateno, Y.; Nishikawa, K. Domain Dislocation: A Change of Core Structure in Periplasmic Binding Proteins in Their Evolutionary History. *J. Mol. Biol.* **1999**, *286* (1), 279–290. https://doi.org/10.1006/JMBI.1998.2454.

(116) Edwards, K. A. Periplasmic-Binding Protein-Based Biosensors and Bioanalytical Assay Platforms: Advances, Considerations, and Strategies for Optimal Utility. *Talanta Open* **2021**, *3*, 100038. https://doi.org/10.1016/J.TALO.2021.100038.

(117) Tam, R.; Saier, M. H. Structural, Functional, and Evolutionary Relationships among Extracellular Solute-Binding Receptors of Bacteria. *Microbiol. Rev.* **1993**, *57* (2), 320–346.

(118) Berntsson, R. P.-A.; Smits, S. H. J.; Schmitt, L.; Slotboom, D.-J.; Poolman, B. A Structural Classification of Substrate-Binding Proteins. *FEBS Lett.* **2010**, *584* (12), 2606–2617. https://doi.org/10.1016/J.FEBSLET.2010.04.043.

(119) Scheepers, G. H.; Lycklama a Nijeholt, J. A.; Poolman, B. An Updated Structural Classification of Substrate-Binding Proteins. *FEBS Lett.* **2016**, *590* (23), 4393–4401. https://doi.org/10.1002/1873-3468.12445.

(120) Pandey, S.; Phale, P. S.; Bhaumik, P. Structural Modulation of a Periplasmic Sugar-Binding Protein Probes into Its Evolutionary Ancestry. *J. Struct. Biol.* **2018**, *204* (3), 498–506. https://doi.org/10.1016/J.JSB.2018.09.006.

(121) Linton, K. J.; Higgins, C. F. The Escherichia Coli ATP-Binding Cassette (ABC) Proteins. *Mol. Microbiol.* **1998**, *28* (1), 5–13. https://doi.org/10.1046/J.1365-2958.1998.00764.X.

(122) Decottignies, A.; Goffeau, A. Complete Inventory of the Yeast ABC Proteins. *Nat. Genet. 1997 152* **1997**, *15* (2), 137–145. https://doi.org/10.1038/ng0297-137.

(123) Garcia, O.; Bouige, P.; Forestier, C.; Dassa, E. Inventory and Comparative Analysis of Rice and Arabidopsis ATP-Binding Cassette (ABC) Systems. *J. Mol. Biol.* **2004**, *343* (1), 249–265. https://doi.org/10.1016/J.JMB.2004.07.093.

(124) Vasiliou, V.; Vasiliou, K.; Nebert, D. W. Human ATP-Binding Cassette (ABC) Transporter Family. *Hum. Genomics* **2009**, *3* (3), 281–290. https://doi.org/10.1186/1479-7364-3-3-281/TABLES/3.

(125) Moitra, K.; Dean, M. Evolution of ABC Transporters by Gene Duplication and Their Role in Human Disease. *Biol. Chem.* **2011**, *392* (1–2), 29–37. https://doi.org/10.1515/BC.2011.006/MACHINEREADABLECITATION/RIS.

(126) Louie, G. V. Porphobilinogen Deaminase and Its Structural Similarity to the Bidomain Binding Proteins. *Curr. Opin. Struct. Biol.* **1993**, *3* (3), 401–408. https://doi.org/10.1016/S0959-440X(05)80113-7.

(127) Michel, F.; Romero-Romero, S.; Höcker, B. Retracing the Evolution of a Modern Periplasmic Binding Protein. *Protein Sci.* **2023**, *32* (11). https://doi.org/10.1002/pro.4793.

(128) Chandravanshi, M.; Tripathi, S. K.; Kanaujia, S. P. An Updated Classification and Mechanistic Insights into Ligand Binding of the Substrate-Binding Proteins. *FEBS Lett.* **2021**, *595* (18), 2395–2409. https://doi.org/10.1002/1873-3468.14174.

(129) Höcker, B.; Schmidt, S.; Sterner, R. A Common Evolutionary Origin of Two Elementary Enzyme Folds. *FEBS Lett.* **2002**, *510* (3), 133–135. https://doi.org/10.1016/S0014-5793(01)03232-X.

(130) Lang, D.; Thoma, R.; Henn-Sax, M.; Sterner, R.; Wilmanns, M. Structural Evidence for Evolution of the β/α Barrel Scaffold by Gene Duplication and Fusion. *Science (80-.).* **2000**, *289* (5484), 1546–1550. https://doi.org/10.1126/SCIENCE.289.5484.1546.

(131) Farías-Rico, J. A.; Schmidt, S.; Höcker, B. Evolutionary Relationship of Two Ancient Protein Superfolds. *Nat. Chem. Biol. 2014 109* **2014**, *10* (9), 710–715. https://doi.org/10.1038/nchembio.1579.

(132) Shanmugaratnam, S.; Eisenbeis, S.; Höcker, B. A Highly Stable Protein Chimera Built from Fragments of Different Folds. *Protein Eng. Des. Sel.* **2012**, *25* (11), 699–703. https://doi.org/10.1093/PROTEIN/GZS074.

(133) Bharat, T. a M.; Eisenbeis, S.; Zeth, K.; Ho, B. A □⌴ -Barrel Built by the Combination of Fragments. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (29), 9942–9947. https://doi.org/10.1073/pnas.0802202105.

(134) Toledo-Patino, S. On the Emergence of the HemD-like Fold and Its Use for Fold-Chimeragenesis, Eberhard Karls Universität Tübingen, 2019.

(135) Braun, A. Protein Design Inspired by Nature: A Study of Cobalamin-Binding Chimeras, University of Bayreuth, 2021.

(136) Szilágyi, A.; Györffy, D.; Závodszky, P. Segment Swapping Aided the Evolution of Enzyme Function: The Case of Uroporphyrinogen III Synthase. *Proteins Struct. Funct. Bioinforma.* **2017**, *85* (1), 46–53. https://doi.org/10.1002/PROT.25190.

(137) Rost, B. Twilight Zone of Protein Sequence Alignments. *Protein Eng. Des. Sel.* **1999**, *12* (2), 85–94. https://doi.org/10.1093/PROTEIN/12.2.85.

(138) Fukami-Kobayashi, K.; Tateno, Y.; Nishikawa, K. Domain Dislocation: A Change of Core Structure in Periplasmic Binding Proteins in Their Evolutionary History. *J. Mol. Biol.* **1999**, *286* (1), 279–290. https://doi.org/10.1006/JMBI.1998.2454.

(139) Bhargav, S. P.; Vahokoski, J.; Kallio, J. P.; Torda, A. E.; Kursula, P.; Kursula, I. Two Independently Folding Units of Plasmodium Profilin Suggest Evolution via Gene Fusion. *Cell. Mol. Life Sci.* **2015**, *72* (21), 4193–4203. https://doi.org/10.1007/S00018-015-1932-0/FIGURES/6.

(140) Almeida, V. M.; Frutuoso, M. A.; Marana, S. R. Search for Independent (β/α)4 Subdomains in a (β/α)8 Barrel β-Glucosidase. *PLoS One* **2018**, *13* (1), e0191282. https://doi.org/10.1371/JOURNAL.PONE.0191282.

(141) Höcker, B.; Beismann-Driemeyer, S.; Hettwer, S.; Lustig, A.; Sterner, R. Dissection of a (Bα)8-Barrel Enzyme into Two Folded Halves. *Nat. Struct. Biol. 2001 81* **2001**, *8* (1), 32–36. https://doi.org/10.1038/83021.

(142) Copley, S. D.; Copley, S. D. Evolution of New Enzymes by Gene Duplication and Divergence. *FEBS J.* **2020**, *287* (7), 1262–1283. https://doi.org/10.1111/FEBS.15299.

(143) Conant, G. C.; Wolfe, K. H. Turning a Hobby into a Job: How Duplicated Genes Find New Functions. *Nat. Rev. Genet. 2008 912* **2008**, *9* (12), 938–950. https://doi.org/10.1038/nrg2482.

(144) Mallik, S.; Tawfik Id, D. S. Determining the Interaction Status and Evolutionary Fate of Duplicated Homomeric Proteins. **2020**. https://doi.org/10.1371/journal.pcbi.1008145.

(145) Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Gould, S. M. Q.; Roodveldt, C.; Tawfik, D. S. The “evolvability” of Promiscuous Protein Functions. *Nat. Genet. 2004 371* **2004**, *37* (1), 73–76. https://doi.org/10.1038/ng1482.

(146) Pandey, S.; Phale, P. S.; Bhaumik, P. Structural Modulation of a Periplasmic Sugar-Binding Protein Probes into Its Evolutionary Ancestry. *J. Struct. Biol.* **2018**, *204* (3), 498–506. https://doi.org/10.1016/J.JSB.2018.09.006.

(147) Gabler, F.; Nam, S. Z.; Till, S.; Mirdita, M.; Steinegger, M.; Söding, J.; Lupas, A. N.; Alva, V. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr. Protoc. Bioinforma.* **2020**, *72* (1), e108. https://doi.org/10.1002/CPBI.108.

(148) Söding, J. Protein Homology Detection by HMM-HMM Comparison. *Bioinformatics* **2005**, *21* (7), 951–960. https://doi.org/10.1093/bioinformatics/bti125.

(149) Bernhardt, H. S. The RNA World Hypothesis: The Worst Theory of the Early Evolution of Life (except for All the Others)A. *Biol. Direct* **2012**, *7*, 1–10. https://doi.org/10.1186/1745-6150-7-23.

12. Acknowledgments

I would like to express my greatest gratitude to Professor Birte Höcker for the opportunity to undertake this thesis with her guidance and support. Her commitment to my intellectual growth and her valuable insights have greatly shaped the direction of my academic life for the better.

I would like to extend my sincere appreciation to all members of the Höcker-lab:

Sergio Romero-Romero, it brought me great joy to have been able to work together with you for all this time. Your input and the discussions, inside the lab and out, made the challenges we faced always seem surmountable.

Sooruban, your technical expertise and professionalism in running the lab really made sure to provide an environment where excellent science could happen, and I appreciate all the input and help you provided over all these years. Especially your expertise on matters of crystallography really helped me reach my goals.

Sabrina, it was always fun to work with you in the lab, your diligence and positive mindset really helped in keeping up the mood even on tricky days.

Francisco, Bruce, Horst, and Abhishek: You really helped shape the direction of my thesis. The discussions we shared in the earlier phase of my time in the group really influenced the way I think about science, and your input was greatly appreciated.

Saacnicteh and Noelia, thank you for your previous work, without which it would have been impossible for me to add to that foundation of insight you provided. The science on the early events of protein evolution – especially the fragments, chimeras and of course Fuzzle – really enabled me to make my own contribution. It really was a blast!

To all my fellow group members, Pascal, Sebastian, Sina, Surbhi, Josef, Merve, Jakob, Andreas, all the Julians, and the many more I had the chance to work with over the years: You were a significant part of what made it so much fun to work in this group. All the discussions, the cooperation, the feedback, and advice really helped create a welcoming and productive atmosphere, and I will miss working with all of you.

Of course, I also owe special thanks to my family, my parents Klaus und Doris, as well as my sister Eva. Your unwavering support even in the more trying times and the never-ending belief that I can pull it off helped me keep on course and actually do it.

Thank you!

(Eidesstattliche) Versicherungen und Erklärungen

(§ 8 Satz 2 Nr. 3 PromO Fakultät)

Hiermit versichere ich eidesstattlich, dass ich die Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe (vgl. Art. 97 Abs. 1 Satz 8 BayHIG).

(§ 8 Satz 2 Nr. 3 PromO Fakultät)

Hiermit erkläre ich, dass ich die Dissertation nicht bereits zur Erlangung eines akademischen Grades eingereicht habe und dass ich nicht bereits diese oder eine gleichartige Doktorprüfung endgültig nicht bestanden habe.

(§ 8 Satz 2 Nr. 4 PromO Fakultät)

Hiermit erkläre ich, dass ich Hilfe von gewerblichen Promotionsberatern bzw. –vermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe noch künftig in Anspruch nehmen werde.

(§ 8 Satz 2 Nr. 7 PromO Fakultät)

Hiermit erkläre ich mein Einverständnis, dass die elektronische Fassung der Dissertation unter Wahrung meiner Urheberrechte und des Datenschutzes einer gesonderten Überprüfung unterzogen werden kann.

(§ 8 Satz 2 Nr. 8 PromO Fakultät)

Hiermit erkläre ich mein Einverständnis, dass bei Verdacht wissenschaftlichen Fehlverhaltens Ermittlungen durch universitätsinterne Organe der wissenschaftlichen Selbstkontrolle stattfinden können.

…….

Ort, Datum, Unterschrift