Short Protein Evolution Stories: Illuminating Aspects of MempromCC and Histidine Kinase Evolution

Dissertation

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

VORGELEGT VON

Ioanna Karamichali
aus Xanthi, Thrakien, Griechenland

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 08.11.2018

Dekan: Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter Prof. Dr. Karl Forchhammer
2. Berichterstatter Prof. Dr. Andrei N. Lupas

"The meaning and purpose of dancing
is the dance"

Alan W. Watts

Table of Content

Abbreviations

6xHis-tag	6 histidine tag aminoacid sequence
B52	DNA gyrase
BDR-repeat	Group of proteins contains the KID repeat as found in *Borrelia*
Buffer A	30mM Tris, 200mM NaCl, 5mM citric acid, 5mM 2-mercaptoethanol, 5μM ZnCl2, 10% glycerol with a pH equal to 8.5
Buffer B	30mM Tris, 500mM NaCl, pH 7-8.5
Buffer C	30mM Tris, 150mM NaCl, pH 8.5
Buffer D	0.1mM Tris, 50mM KCl, 5mM β-mercaptoethanol, 5% glycerol
CA	Catalytic-ATP binding domain
CCDC90A	Same as MCU1 protein, MempromCC family member that binds MCU and Ca2+ in Humans
CCDC90B	Homolog of MCU1 protein, MempromCC family member
CheA	E. coli chemotaxis motility TCST system
Cl-LVA	LVA destabilized Lambda cI repressor
Cph1	E. coli photosensor TCST system
Cph1/EnvZ	Chimeric photosensor TCST system
Dali	Online structure comparison tool
DHp	Dimerization and histidine phosphotransfer domain
DHp-CA	Homodimeric transmitter domain of EnvZ TCST system which is found in the cytosol and includs the DHp and CA domains
DHp-CA(T19Q)	Threonine-Glutamine mutant on the 19th amino acid of the DHp-CA EnvZ transmitter
DHp-DX1	Fusion protein with natural EnvZ linker
DHp-DX1(H15Q)	DHp-DX1histidine defficien mutant on the amino acid in possition 15
DHp-DX2	Fusion protein with designed G-rich linker
DnaK/J	Heat shock proteins induced by 2% ethanol
DX	In vitro evolved ATP binding protein, also known as 18-19, FOB, FamB, ANBP in different stages of in vitro evolution
DX-DHp	Fusion model with inversed domain order
EnvZ	Osmosensing TCST system
GFP	Green fluorescent protein
gfp_LVA	LVA destabilized version of the green fluorescent protein
GHKL	Bacterial gyrase, HSP90, histidine kinase and MutL superfamily that contains homologs of the CA domain
GroES/EL	Heat shock proteins induced by 2% ethanol
HAMP	Motion-couple domain present in Histidine kinases, Adenyl cyclases, Methyl-accepting proteins and Phosphatases
HAMPAf1503	HAMP domain found in a sensor-HAMP only system of Archaeoglobus fulgidus
HNSA	Head-Neck-Stack-Anchor secondary structure
HPt	Histidine phosphorylation
Hsp90	ATPase with CA protein homologs
IEF	Isoelectric focusing analysis
IP	Immunoprecipitation

LVA	Protease recognised destabilizing peptide tail with the sequence "AANDENYALVA"
MALS	Multiangle light scattering
MCU	Mitochondrial Ca^{2+} uniporter in Humans
MCUR1	Same as CCDC90A protein, MempromCC family member that binds MCU and Ca2+ in Humans
MST	Microscale thermophoresis
MutL53	DNA mismatch repair protein
N,G1,F,G2	Conserved amino-acid boxes characteristic for the CA domain
NMR	Nuclear magnetic resonance
ompC	OmpC porin promoter activated mostly under lower osmolarity stresses through the phosphorylated OmpR response regulator
ompF	OmpF porin promoter activated mostly under higher osmolarity stresses through the phosphorylated OmpR response regulator
OmpR	Response regulator of the EnvZ TCST system
P_L	Bacteriophage lambda promoter
psaA	Protein component of the Photosystem I
Rat1	MempromCC family member found in chloroplasts and an alternative maturation control factor of the PsaA protein
RR	Response regulator
SacB	Gene toxic in the presence of sucrose
SEC	Size exclusion Chromatography
SpoIIAB	CA homodimeric serine kinases in *Bacillus*
T-DHp-DX1	Fusion protein of Tar sensor, HAMP domain and the DHp-DX1protein
Tar	Aspertate chemoreceptor
Taz	Chimeric two component signal transduction system that combines the Tar chemoreceptor with the cytosolic part of the EnvZ osmosensor.
TCST	Two component signal transduction
TEV	Protease
TRX	Thioredoxin fusion that facilitates the stabilization of proteins
tscA	RNA cofactor
YpdIp	HPt containing protein that does not include the CA domain in yeast
λred	Lambda red recombinase

Zusammenfassung

Sequenz-basierte bioinformatische Analysen waren lange die treibende Kraft in Evolutionsstudien, obwohl auch experimentelle Vorgehensweisen durch Mutagenese, *in vivo* und *in vitro* Evolution neue Einblicke liefern. In dieser Arbeit wurden beide Vorgehensweisen angewendet, um evolutionäre Aspekte zweier verschiedener Protein-Familien, der mempromCC-Proteine und der Histidinkinasen, genauer zu beleuchten.

Die mempromCC-Familie besteht aus membrangebundenen coiled-coil-haltigen Proteinen, welche hauptsächlich in den Mitochondrien zu finden sind. Unsere bioinformatischen Analysen verknüpfen diese Proteine zum ersten Mal und ermöglichen eine gemeinsame strukturelle und funktionelle Analyse der mempromCC-Familie. Obwohl ihr Funktionsmechanismus weiter ungeklärt bleibt, zeigte sich, dass diese Proteine als Assemblierungsfaktoren für verschiedene Proteine dienen, zum Beispiel den mitochondrialen Ca^{2+} Uniporter (MCU), Cytochrom C und Photosystem I, welche wichtig für das Überleben der Zellen sind.

Histidinkinasen wirken typischerweise als Multidomänen-Proteine, bestehend aus Transmembran-Sensor- und cytoplasmatischen Effektor-Domänen. In der osmo-regulierenden Histidinkinase EnvZ werden Signale durch die DHp (Dimerisierungs- und Histidin-Phosphorylierungs-domäne) und die CA (katalytische- und ATP-bindende Domäne) Domäne übertragen: CA trägt ATP und phosphoryliert ein Histidin in DHp, welches die Phosphatgruppe dann auf nachgeschaltete Effektoren überträgt. Dies führt in Genen, welche vom OmpC-Promotor reguliert werden, zu einer Transkriptionsänderung. Basierend auf den evolutionären Merkmalen, die sowohl für DHp als auch CA beobachtet wurden, ist ein Evolutions-Szenario entstanden. Die beiden Domänen wurden früh in der Evolution fusioniert und bildeten aus einem einfachen ATP-bindenden Element eine Histidinkinase. Um diese Möglichkeit zu untersuchen, haben wir eine Chimäre hergestellt, in welcher CA durch DX ersetzt wurde. DX ist ein künstliches Protein, das via *in vivo*-Evolution durch Selektion auf ATP-Bindeaffinität generiert wurde. Ein DHp-DX-Fusionsprotein (DHp-DX1) zeigte *in vitro* tatsächlich eine starke ATPase-Aktivität, für DX allein wurde dies nicht beobachtet. *In vivo* erhöhte DHp-DX1 die Transkription von OmpC-regulierten Genen im Vergleich zu einer Phosphotransferase-defizienten DHp-DX1(H15Q)-Mutante. Zusammenfassend unterstützen diese Entdeckungen die modulare Entwicklung des DHp-CA-Elements und bieten einen Beweis für das Konzept der Ur-Enzymentwicklung.

Summary

Sequence-based bioinformatic analysis has long been a driving force of evolutionary studies, while experimental approaches offer new insights through mutagenesis, *in vivo* and *in vitro* evolution. Here, we employ both approaches to illuminate aspects of the evolution of two different protein families, the mempromCC protein family and the histidine kinase protein family.

The mempromCC family comprises membrane-bound coiled-coil-containing proteins found mostly in mitochondria. Our bioinformatic analysis connects for the first time these proteins allowing their collective structural and functional analysis. While the mechanism of function remains elusive, these proteins have been shown to act as assembling factors of different proteins like the mitochondrial Ca^{2+} uniporter (MCU), cytochrome c and Photosystem I that are crucial for cell survival.

Histidine kinase proteins typically function as multidomain proteins, comprising transmembrane sensor and cytoplasmic effector domains. In the osmo-regulating histidine kinase EnvZ, signals are transmitted through the DHp (dimerization and histidine phosphotransfer) and the CA (catalytic and ATP-binding) domains: CA carries ATP and phosphorylates a histidine in DHp, which then transfers the phosphate group to downstream effectors, resulting in a modulated transcription of genes controlled by the ompC promoter. Based on the evolutionary traits observed for both DHp and CA an evolutionary scenario has emerged, in which the two domains were fused early in evolution building a histidine kinase from a simple ATP-binding element. To study this possibility, we produced a chimera, in which CA was replaced with DX, an artificial protein that was generated through in-vivo evolution by selecting for ATP binding affinity. Indeed, a DHp-DX fusion protein (DHp-DX1) showed a strong ATPase activity *in vitro*, that was not seen for DX alone. *In vivo*, DHp-DX1 increased the transcription of ompC-controlled genes compared to a phosphotransferase-deficient DHp-DX1(H15Q) mutant. Collectively, these findings support the modular evolution of the DHp-CA element and offer a proof of concept for primordial enzyme evolution.

List of publications

1. Hartmann MD, Mendler CT, Bassler J, **Karamichali I**, Ridderbusch O, Lupas AN, Hernandez Alvarez B. (2016) α/β coiled coils. Elife (doi: 10.7554/eLife.11861). (published)
2. Adlakha J, **Karamichali I**, Sangwallek J, Deiss S, Bär K, Coles M, Hartmann MD, Lupas AN, Hernandez Alvarez B. MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles. (Under submission)
3. **Karamichali I**, Sepulveda E, Zhu H, Coles M, Lupas AN. Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase. (Ready for submission)

Introduction

Many people would describe evolution as the process of improving something, whether this is technology, medicine, research, or society itself. However, in biology evolution is more accurately described as the process of change that is driven by natural selection. The concept of improvement is more complicated in biology because perfectionism does not exist in biology and imperfection is the motor of life itself. Small imperfections constantly accumulate in nature allowing the coexistence of small variations of a characteristic or even the arising of new characteristics that can be selected through generations because they provide an advantage under specific conditions. Life is a struggle for existence within a constantly changing environment that challenges organisms not only to survive but also to compete with each other. A constant struggle that allows the fittest organisms to proliferate their characteristics more successfully under specific conditions, a concept described today as natural selection.

The study of evolution is and always has been of great importance. Even though the concepts of evolution and the questions revolving around it might seem abstract or philosophical to the general population, evolutionary studies have significantly aided developments in many everyday aspects of life. Medicine, for example, has enormously benefited through the understanding of inheritance patterns of disease, or through the development of new drugs and treatments by targeting mechanisms only found in infectious agents like the synthesis of the peptidoglycan cell wall in bacteria. Evolution has also allowed a better understanding of ecosystems aiding decision making regarding their protection, while it has constantly promoted the development of agriculture, farming and biotechnology.

Also of great significance, are the philosophical and social aspects of evolution. The understanding of how the human kind has evolved provided us with the notion that we are a part of nature and responsible for the wellbeing of it as a whole. This notion has transformed and keeps improving our society by directing it towards a more knowledge based, responsible and ethical practice of life. A great example of biological evolution effects on society is the so-called "panda" trial in Dover, Pennsylvania that took place in 2004. During this trial scientists were asked to defend the pupil's right to religious freedom and accurate scientific education by using evolution studies as a weapon. These scientists helped to protect the independence of the educational system and also turned the local politics towards people that valued science. There were, however, cases were evolution lost the battle. An example is the "Monkey" trial that took place in Tennessee in 1925, where a school teacher was convicted and had to pay a 100$ (1440$ in present dollars) fine because he was teaching evolution. As a result, the course of evolution remained absent from the curriculum of schools for another 30 years, setting back the educational development of a whole generation[1,2].

It is more than understandable that there is controversy when it comes to evolution. Darwin himself recognized the difficulty of understanding how evolution can result in complicated and sophisticated biological machinery like the eye as he expressed in a letter to botanist Asa Gray in 1860, saying "The eye to this day gives me a cold shudder, but when I think of the fine known gradation my reason tells me I ought to conquer the odd shudder."[3]. Darwin's reason has since been vindicated because the formation of the eye has been found to evolve independently multiple times during the course of evolution, in a process known as convergent evolution. In Richard Dawkins' words "It has been authoritatively estimated that eyes have evolved no fewer than forty times, and probably more than sixty times, independently in various parts of the animal kingdom. In some cases these eyes use radically different principles. Nine distinct principles have been recognized among the forty to sixty independently evolved eyes."[4].

Eyesight is a highly important sensory mechanism for survival, providing a huge edge to the organisms that possess it. However, higher organisms are not the only ones that benefit from a light sensory system like eyesight. Cyanobacteria,

for example, have multiple photosensors that even detect different light wavelengths. One of these molecules is Cph1, a histidine kinase that senses light in the far red and is probably an ancestor to a photoreceptor found in plants, called Phytochrome I, that acts as a serine/threonine kinase instead[5–7]. This type of evolution is known as divergent evolution, during which a gene is altered to the point of presenting a slightly different or sometimes even a radically different function from the one of its ancestor.

Protein evolution is full of small amazing evolutionary stories. Stories like the independent evolution of analogous-ice-binding-antifreeze proteins through convergent evolution[8], or famous stories like the divergent evolution of the bacterial flagella proton-motor from injectisomes, a group of protein toxins forming a molecular syringe found in many pathogenic bacteria[9]. However, no matter how different evolutionary stories can be, all proteins seem to have started in a common way within a completely foreign RNA world.

The RNA past of our world seems more plausible as increasing number of studies provide evidence regarding the ability of RNA to store genetic information and to catalyze chemical reactions, even in the organisms of today[10–12]. However, though RNA possesses the ability to start, store and replicate the knowhow of catalytic chemistry, the formation of the first cells would be next to impossible without outsourcing of information storage to the more stable and reliable DNA molecules, while chemical catalysis was enhanced by the more flexible and dynamic protein molecules. The critical step of the later transition came with the development of the ability of RNA to bind amino acids and to catalyze peptide bonds, a function RNA still retains today as the main catalyst in the ribosome[13–17].

The first random polypeptides started as simple protein folds that later formed simple protein domains characterized by crude, unspecific multifunctionality. As more flexible generalists these domains were randomly binding ligands and each other, collaborating into functional novelty that could be stored in a newly formed genome as domain fusions, creating the first primordial enzymes[18,19]. The combination of domains lead to more sophisticated catalytic machines directing enzyme functionality towards a specific function, but retaining the nature of a generalist in each single domain[20–22]. This enzyme multifunctionality,

as understood today, allows divergent evolution driven by different evolutionary trajectories to specialize enzymes towards different functionalities[23–25], while domain reshuffling redistributes the functional vocabulary[26,27]. Primordial enzymes like modern enzymes could show a completely different or a related functionality to their descendants[23].

In this study I will attempt to decipher basic principles of protein evolution by investigating the evolutionary history of two protein families, the recently discovered protein family mempromCC and the in depth studied histidine kinase protein family. This trip back in time will showcase two very different approaches of studying protein evolution. First by simply collecting and observing the protein "fossils" in the sequence database that is available today. And second, by actively manipulating the process of evolution through "intelligent design" in order to replicate or test different evolutionary theories.

Evolutionary story 1: MempromCC proteins

Coiled-coils are protein bundles of two or more α-helices twisted around each other. They are characterized by specific coiled-coil pitch periodicities of usually 7, 11, or 15 residues, forming 2, 3 or 4 α-helical turns, respectively. Coiled-coils can be detected based on their amino acid periodicity, while their backbone structure can be accurately predicted based on parametric equations[28,29]. These protein structures can be associated with many protein functions because they can form protein fibers and motor proteins but also barrels, funnels, sheets and spirals[30,31]. Many proteins have been found to contain coiled-coils among different secondary structures, some of which can be found to extend from a membrane using a coiled-coil stalk as a molecular spacer[29,31].

We have recently found a group of coiled-coil-membrane proteins that are uniquely characterized by one or multiple 6 residue-long β-strand necks placed between an N-terminal α-helical head and a coiled-coil stalk (Figure 1). These proteins can be found all over the tree of life, and even though they vary greatly in size and sequence, their stable head-neck-stalk-anchor (HNSA) secondary

structure, along with sequence similarities found between the necks and other head sequence patterns, raised the hypothesis of an evolutionary common root.

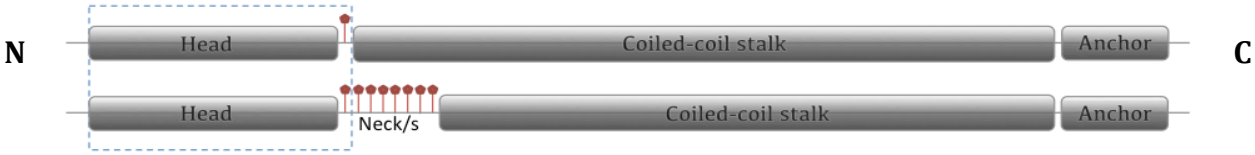


Figure 1. Cartoon representing the HNSA secondary structure of mempromCC proteins. The dotted lines indicate the head-neck portion used to search and collect more proteins with the same secondary structure. The same portion was used for the final family classification.

This hypothesis was tested to verify whether a currently unknown coiled-coil containing membrane protein family, here called mempromCC, could be identified and further characterized. To investigate the evolutionary history of these proteins, as well as their possible structure and function, I used a number of head-neck sequences to collect more proteins with the same characteristics. 1085 proteins were identified based on deep evolutionary searches, multiple sequence alignments and secondary structure prediction. These 1085 proteins were classified based on their sequence similarities, aiming to reveal any evolutionary connections between them, while in parallel a possible common structure and function was explored based on previous studies involving the detected proteins.

My study of the HNSA proteins revealed that many of them are indeed related forming the new mempromCC protein family. MempromCC proteins can be found in both bacteria and eukaryotic organelles and they show a characteristic and highly conserved head-structural-core. While their function remains unclear there are significant functional differences between different homologs making mempromCC protein evolution an example of divergent evolution.

Evolutionary story 2: Histidine kinases

Histidine kinases transfer a phosphate group from ATP to a histidine residue within the kinase. They are abundant in bacteria where they prevail as the main environment sensing system[32]. There are two types of histidine kinases, type I and II, which differentiate in their domain variation and organization[33]. The type I histidine kinases have the combined DHp-CA domains (known as transmitter) as their cytosolic-main catalytic part. During evolution more domains were gradually added to the transmitter like a sensor domain that connects the system to the external environment allowing the sensing of a variety of signals[33–35]. An example of such a system is the EnvZ two component signal transduction (TCST) system, that has been extensively studied[36–38] (Figure 2.A.).

EnvZ is a type I histidine kinase that regulates the expression of proteins that form pores through the cell membrane, the so-called porins, that manage the osmotic or chemical stress[39–41]. The osmotic alterations are sensed through the EnvZ sensor that transfers the signal through the linker domain HAMP (present in Histidine kinases, Adenylate cyclases, Methyl-accepting proteins and Phosphatases) resulting in the autophosphorylation of the cytosolic transmitter on the histidine found on DHp and the sequential phosphorylation of the response regulator OmpR on an aspartate phosphoreceptor[36].

The sensing domains from TCST systems can vary greatly indicating that different domains have been fused to the system during evolution[33]. This plasticity of TCST systems has allowed the creation of chimeras like the Tar/EnvZ (Taz chimera) or the Cph1/EnvZ created by exchanging the osmosensor of EnvZ with either the aspartate sensor Tar[42] (Fig 2B) or the photosensor Cph1[43] (Fig 2C). Both of these systems have been used to study TCST systems *in vivo*. This was possible after coupling the chimeric histidine kinase to the expression of a green fluorescence protein (GFP) which was directly or indirectly controlled by the ompC promoter triggered by the phosphorylated response regulator OmpR (Figure 2.B. and C.)[44–46].

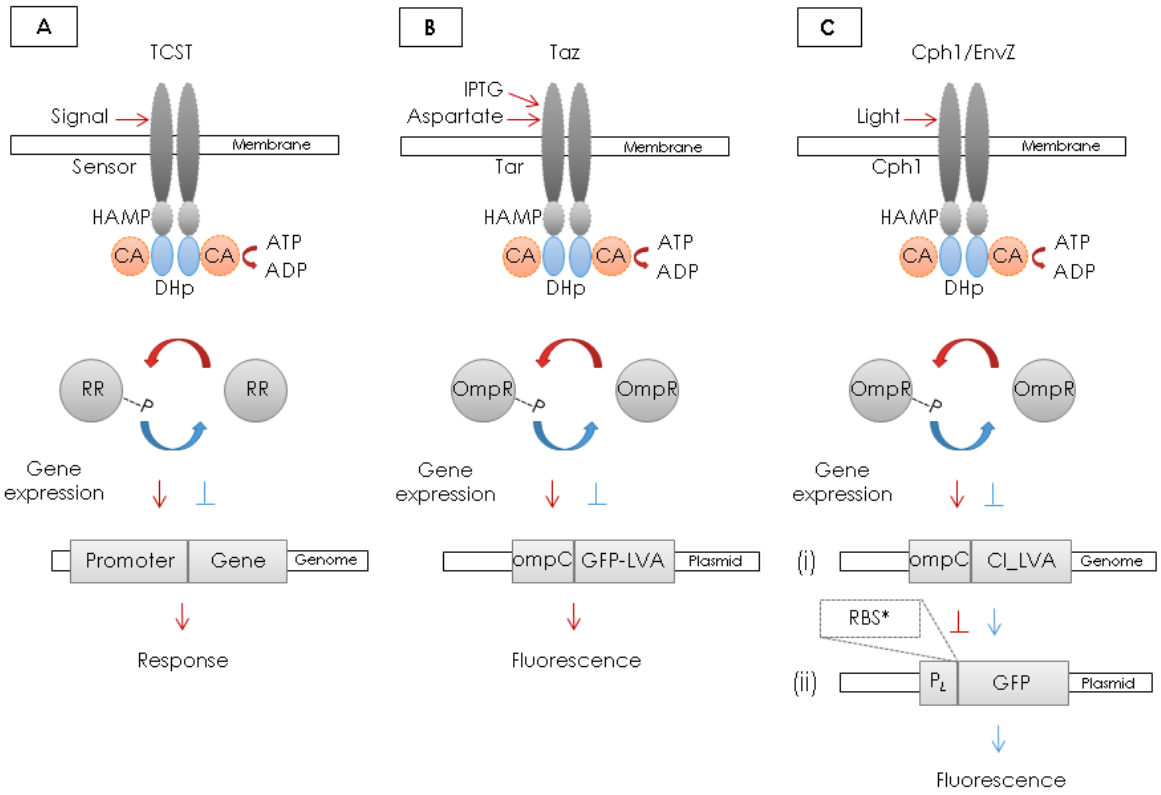


Figure 2. Cartoon representation of two component signal transduction (TCST) systems. (A) General representation of the most common TCST systems. The sensor reacts to a signal, the stimulus is transferred through the HAMP domain to the DHp-CA transmitter, which is autophosphorylated in response. Finally, the transmitter reacts with the response regulator (RR) through the DHp domain, leading to the phosphorylation of the RR on an aspartate residue. The system usually also has a dephosphorylation activity when a signal is absent. The phosphorylated RR can directly or indirectly trigger gene expression through a promoter. **(B)** The chimeric Taz system developed by Michalodimitrakis *et al.* (2005), which combines the Tar chemoreceptor with the cytosolic part of the EnvZ osmosensing TCST system. An aspartate signal triggers the autophosphorylation of the transmitter, resulting in the phosphorylation of the OmpR response regulator, which leads to the expression of the GFP-LVA gene through the ompC promoter. **(C)** The light-switchable TCST system (Cph1/EnvZ) developed by Lee *et al.* (2013), which combines the Cph1 light sensor with the cytosolic part of the EnvZ osmosensing TCST system. In this case the ompC promoter is placed on the chromosome instead of a plasmid. There is also a two step control of reporter gene expression. (i) ompC controls the expression of the repressor CI-LVA. (ii) CI-LVA represses the expression of GFP through the highly controlled second promoter (P_L). Here both a weak and a strong RBS can be used (*). Additionally, the protease recognized 11-amino-acid long LVA tail is present in reporter proteins of both systems, increasing their sensitivity.

The ATP binding domain CA is part of the transmitter and presents an interesting evolutionary history. The homology of the CA domain is determined by the highly conserved N,G1,F,G2 amino-acid boxes and the shared 4 α-helices-5 β-strands sandwich structural architecture, known also as the Bergerat fold (Figure 3.A.)[33,47,48]. The CA domain of EnvZ shares the same characteristics but differs in some loops that differ between CA homologs, like the CheA and SpoIIAB CA homologs that have extra α-helices[33,49]. Figure 3A represents the EnvZ structural variation in gray, while the blue areas show the conserved Bergerat fold. The long loop between the α3 and α5 α-helix plays an important role since it acts as a lid protecting the ATP binding pocket that is found between the β-sheets β5-7 and the α-helices α2, α3 and α5, an area bordered by all four conserved boxes.

The CA domain is present and highly conserved in both types of histidine kinases, where it is responsible for binding the ATP that is hydrolysed in response to a signal, leading to the transfer of the γ-phosphate onto a histidine residue found either on the DHp (Dimerization and histidine phosphorylation) domain for the type I histidine kinases, or on the HPt (Histidine phosphorylation) domain for the type II histidine kinases. Even though both DHp and HPt domains share a very similar 4 helix bundle fold, their similarities are superficial since they present no sequence similarity and they have very different functionality and evolutionary history. HPt has been found independent from the CA domain in yeast (YpdIp) while DHp has been found only in combination with the CA domain, a fact that indicates that CA and DHp have a long history of coevolution, being fused very early in their existence. Interestingly, *in vitro* separated DHp domains can also present phosphatase activity while this is not the case for HPt domains[33].

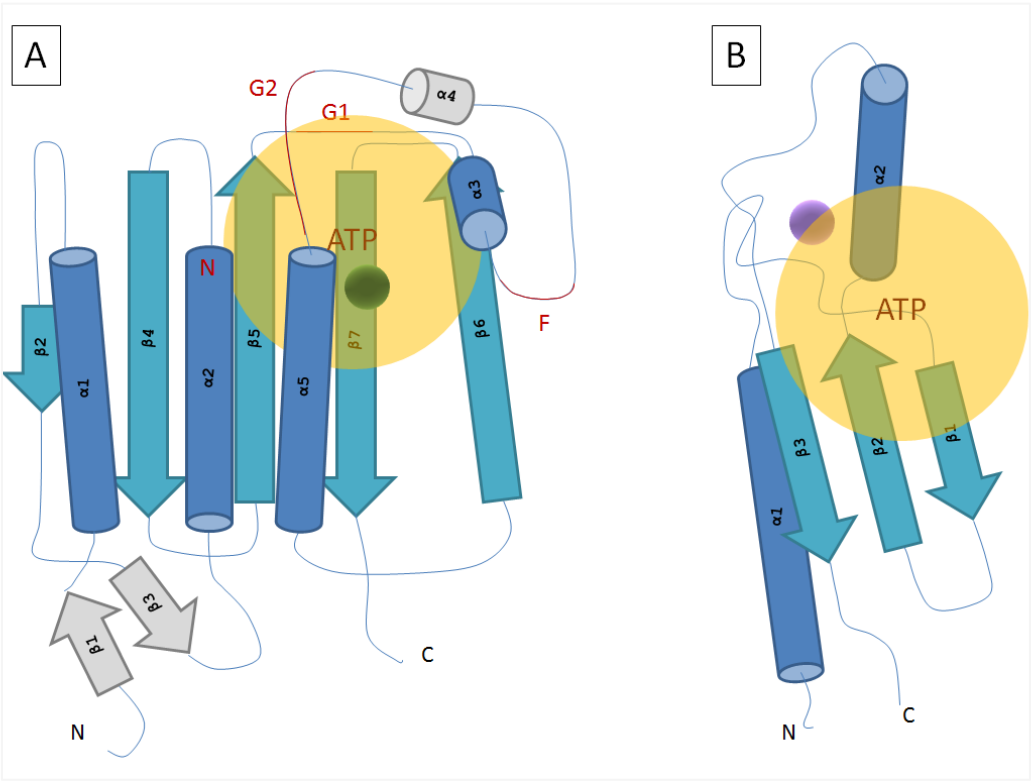


Figure 3. Cartoon representation of the CA and DX protein structures. (A) CA domain structural representation. The gray secondary structure elements are unique to the EnvZ CA domain, the blue ones are conserved structural elements found in the Bergerat fold. The conserved sequence boxes N, G1, F and G2 are shown in a red color.**(B)** DX domain structural representation. The protein forms a zinc finger. The ATP is quite exposed and bound between α-helix α2 and the β-sheets β1-β3. No Mg^{2+} was found bound to the DX structure, while the ATP formed a conformation with an unusual bend. The approximate position of the ATP binding pocket is presented as orange shadow. Mg^{2+} and Zn^{2+} metals are represented as green and purple spheres, respectively.

The CA domain demonstrates an interesting evolutionary history that in many cases differs from the one of the DHp domain or the histidine kinases, offering a great divergent evolution case. Homologs of the CA domain can be found as independent proteins that function as homodimeric serine kinases called SpoIIAB in *Bacillus* species[6,49,50]. Other reports also indicate homologs of CA to be part of the two component like-mitochondrial serine kinases called BCKD[51,52], but also ATPases like Hsp90[53], the DNA gyrase B[54] and the DNA mismatch repair protein MutL[55], which are members of the GHKL (bacterial gyrase, HSP90, histidine kinase, MutL) superfamily. These findings lead to the hypothesis that the CA domain was evolved to bind ATP and later was combined with different

domains to evolve a more complicated functionality that can include ATP hydrolysis, and phosphorylation of either histidine or serine amino acids of other domains.

Another example of an interesting evolution story is the case of the novel ATP-binding protein DX. This protein was *de novo* evolved *in vitro* after multiple cycles of mutagenesis and selection based on the ability to bind ATP. The process of evolution was based on a starting library (6×10^{12}) of random-80 amino acid long peptides, random mutagenesis and RNA display[56]. DX is a zinc finger protein that forms an α-helix-three β-sheets-α-helix fold (Figure 3.B.). The ATP binding site is a hydrophobic pocket found between the three β-sheets and helix 2, which are connected by a rather long loop that borders the binding site. The binding pocket is rather shallow leaving exposed the sugar and phosphate moieties of the ATP molecule[57,58]. DX seems to bind both ATP and ADP with ATP being the dominant ligand favored by up to 2.5-fold [57]. ATP has been reported to adopt an unusual bended conformation, which has been hypothesized to assist metal independent-slow hydrolysis of ATP since no Mg^{2+} was found bound to the structure, a hypothesis that has remained unverified.

The interesting independent evolutionary history of CA along with the achievement of *de novo* evolving a protein like DX offers the opportunity to replicate the evolution of a newly formed ATP binding domain into a functionally and structurally more sophisticated primordial enzyme. Here I will demonstrate this replication by evolving a primordial enzyme precursor of a histidine kinase through the fusion of the DX ATP binding domain to the EnvZ DHp domain. This demonstration offers a proof of concept of the evolution of primordial enzymes, while showing a case where a true analog replaces a protein domain within a functional system *in vivo.* This could show that the selection of the CA domain that constitutes part of the crucial for bacterial survival TCST systems was selected based on chance and opportunity, since another ATP binding domain could possibly serve the same purpose. Such enzyme functional evolution could also significantly aid the challenging enzyme design process by exploiting domain recycling, *in vitro-de novo* protein evolution and *in vivo* selection as a new approach in creating novel enzymes.

Aims and significance

In this study two approaches were used to answer fundamental questions about the evolutionary history of two protein families, the mempromCC and the histidine kinase protein family.

Firstly, the new protein family mempromCC was identified and classified using bioinformatic-secondary structure and sequence analysis. This sequence based evolutionary analysis shows how evolutionary studies can link distant homologs and seemingly unrelated studies, aiding the structural and functional characterization of proteins.

Secondly, a new approach to testing the functional evolution of primordial enzymes recreated the evolution of a primordial histidine kinase. The *in vitro* evolved ATP binding protein DX was fused to DHp, a domain found in the transmitter of the EnvZ histidine kinase, aiming to replace the natural ATP binding domain CA. This approach is significant on multiple levels: it answers fundamental questions regarding the evolution of primordial enzymes, it replicates the possible evolutionary history of histidine kinases, while it also demonstrates that analogous proteins, products of convergent evolution, can replace each other into a functional system in the cell. Finally, our approach of combining *in vitro* evolution, natural domain recycling and *in vivo* evolution could aid enzyme design, leading to the creation of novel enzymes.

MempromCC: A new family of coiled-coil containing proteins

Project Summary

We have recently discovered a protein group that is characterized by a unique secondary structure of an α-helical head, a β-layer neck, a coiled-coil stalk and a trans-membrane anchor (HNSA; Figure 1). I have detected and collected protein sequences with a predicted HNSA secondary structure based on the homology of the helical head and the presence of a neck between the head and the coiled-coil stalk.

The collected HNSA dataset was used as a set of proteins that contain folds of coiled coils alternating with β-strands, or α/β fibers as described in the enclosed paper "α/β coiled coils"[29] (see Appendix). The paper argues that the insertion of 2 or 6 amino-acids in the coiled coil heptad periodicity can increase the pressure of the fold to a breaking point, leading to the formation of short β-strands that turn the chain by 120°. My HNSA protein dataset contained examples of this fiber formation in both prokaryotes and eukaryotic organelles, underlining the significance of this new finding.

I further classified the HNSA protein dataset using the head-neck portion of the proteins that showed the highest conservation rate (Figure 1). My analysis indicates that even though the neck appears to be strikingly similar among the vast majority of sequences and the α-helical head is conserved in a great variety of organisms, not all HNSA proteins seem to be evolutionary connected. Support for this finding is the structure variation within the group and the great sequence variation of the proteins.

However, a conserved family of proteins can be distinguished within the HNSA group, the so-called mempromCC protein family. MempromCC proteins are mostly found in α-, β-, γ- and δ-proteobacteria, eukaryotic organelles of fungi,

metazoa, plants and some algae, as well as in some groups of *Pseudomonas*, *Delftia* and *Chromatiales*. The strong conservation observed for the core of the helical heads of this family, along with a structural analysis of one of the family representatives enabled the structural representation of more members of this family. The complete bioinformatics analysis is presented in the enclosed paper "MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles" (see Appendix). This paper focuses on the structural and functional analysis of one of the mempromCC representatives, called 90A or MCUR1, arguing that this protein interacts directly with the mitochondrial calcium uniporter (MCU) through the head domain.

Most MempromCC proteins have not been functionally characterized and are often annotated as hypothetical, however MCUR1[59,60], and Rat1[61] have been previously studied. Interestingly, both seem to play a role in the assembly of two different components of the electron transport chain in eukaryotes, the cytochrome c and the photosystem I. Rat 1 has also been reported to bind RNA. The association of both proteins with the assembly of members of the electron transport chain , as well as the direct connection of Rat1 to the splicing mechanism of Group II introns, raise the possibility of an evolutionary and functional connection between mempromCC proteins and Group II introns[62]. The lack of a knock out phenotype along with the coexistence of mempromCC-like proteins and Group II introns in bacteria, archaea, mitochondria and chloroplasts provide further support for this hypothesis.

α/β coiled coils

Hartmann MD, Mendler CT, Bassler J, **Karamichali I**, Ridderbusch O, Lupas AN, Hernandez Alvarez B

Elife (doi: 10.7554/eLife.11861).

2016

(see Appendix)

Abstract

Coiled coils are the best-understood protein fold, as their backbone structure can uniquely be described by parametric equations. This level of understanding has allowed their manipulation in unprecedented detail. They do not seem a likely source of surprises, yet we describe here the unexpected formation of a new type of fiber by the simple insertion of two or six residues into the underlying heptad repeat of a parallel, trimeric coiled coil. These insertions strain the supercoil to the breaking point, causing the local formation of short β-strands, which move the path of the chain by 120° around the trimer axis. The result is an α/β coiled coil, which retains only one backbone hydrogen bond per repeat unit from the parent coiled coil. Our results show that a substantially novel backbone structure is possible within the allowed regions of the Ramachandran space with only minor mutations to a known fold.

Contribution

I have collected proteins with an α/β coiled coil fold during the analysis and classification of the proteins with a predicted HNSA secondary structure. My contribution to this paper is the sharing of my protein dataset, and participating in the β-layer sequence analysis by extracting and representing the β-layers found, along with their helical flanking regions, as shown in figure 6. The other authors contributed as indicated in the contribution section of the paper[29].

MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles

Adlakha J, **Karamichali I**, Sangwallek J, Deiss S, Bär K, Coles M, Hartmann MD, Lupas AN, Hernandez Alvarez B.
(Submitted; see Appendix)

Abstract

Membrane-bound coiled-coil proteins are important mediators of signaling, fusion, and scaffolding. Here, we delineate a heterogeneous group of trimeric membrane-anchored proteins in prokaryotes and eukaryotic organelles. They exhibit a characteristic head-neck-stalk-anchor architecture, in which a membrane-anchored coiled-coil stalk projects different N-terminal head domains via a β-layer neck. Based on sequence analysis, we identify different types of head domains and determine crystals structures of two representatives, the archaeal protein Kcr-0859 and the human CCDC90B, which possesses the most widespread head type. Using the functionally characterized paralog of CCDC90B, the mitochondrial calcium uniporter regulator 1 (MCUR1), we study the role of individual domains, and find that the head interacts directly with the mitochondrial calcium uniporter (MCU) and is destabilized upon Ca^{2+} binding. Our data provide structural details of a class of membrane-bound coiled-coil proteins and identify the conserved head domain of the most widespread type as a mediator of their function.

Contribution

I contributed to this paper by conducting the bioinformatics analysis via collecting the HNSA protein sequences, analyzing their secondary structure and sequence, and finally classifying them. The analysis started with an initial dataset of proteins provided by Prof. Dr. Andrei Lupas whose input and guidance was crucial for completion of this project. Part of my contribution was the structural modeling of the *Caulobacter* proteins Na1000 and JGI0001013-D04 that were based on the crystal structure of CCDC90B, which was provided by my coworkers and presented in the same paper. I actively participated in the writing, especially for the parts on my work. I was the main contributor of figure 1 and 2 while I also significantly contributed in the preparation of figure 7.

The other contributions are as follows:

1. Jyoti Adlakha: Cloning, protein purification and crystallization setup for CCDC90B-GCN4, CCDC90B-head and MCUR1-head; biophysical characterization (MST, CD, SEC MALS and EM of beta-fibrils); *Caulobacter* cell fractionation; MCUR1 model, purification of labeled protein for NMR
2. Sangwallek Juthaporn: EM of *Caulobacter*
3. Silvie Deiss: *Candidatus Korarchaeum* cloning and purification
4. Marcus D. Hartmann: Structures of *Candidatus* protein and CCDC90B
5. Murray Coles: NMR analysis of MCUR1 head
6. Andrei Lupas: Bioinformatics, initiation of project
7. Birte Hernandez Alvarez: IP experiments in cells, project management

Additional Results/Discussion

Details regarding the size variation and genus distribution of HNSA proteins

The identified 1085 HNSA proteins show great variation in their sequence and size. The total length of the protein sequence usually varied between 60-320 amino-acids, while the heads could be between 40-220 amino-acids long. The longest sequences were mostly *Ascomycota* and *Basidiomycota* proteins while the shortest ones belong to the bacterial genus *Xylella*. Protein sequences with lengths longer or shorter than the ranges above were usually either partial proteins, or proteins with possibly misannotated sequences, a common finding for hypothetical proteins.

HNSA proteins were found in all three domains of life as well as in 5 viruses as shown in table 1. The viral proteins were all found in phages and their sequences were quite similar. These proteins are probably products of a horizontal gene transfer and they play no role in the phage infection cycle, since they are rare. Interestingly, the head-neck sequence of the viral proteins lead to the detection of proteins found in the bacterial genus *Borrelia*, which belong to a family known as the BDR-repeat protein family. The BDR-repeat proteins seem to share the same head-stalk-anchor structure, however the detection of the neck was not possible based only on their sequence, leading to the exclusion of these proteins from the final dataset. All HNSA proteins found have their genes located in the chromosome. Interestingly though, 39% of the eukaryotic proteins have been annotated as mitochondria located proteins while one protein has been located in chloroplasts (Rat1).

Table 1. Distribution of the HNSA protein family. All three domains of life are represented.

Kingdom	Number of sequences	Genus variety
Archaea	35	11
Bacteria	441	105
Fungi	305	13
Plantae	102	48
Animalia	197	93

*Viruses, 5 Phages

MempromCC functional and evolutionary theories

Even though the proteins found were not previously functionally characterized and are mostly annotated as hypothetical, two of them named MCUR1 (*Homo sapiens,* NP_001026883.1, also known as CCDC90A), and Rat1 (*Chlamydomonas reinhardtii*, XP_001694431.1) have been previously studied. MCUR1 (CCDC90A) has been the subject of a long debate since it has been involved in both the regulation of the mitochondrial calcium uniporter (MCU; see enclosed paper "MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles")[59,60] and in the assembly of cytochrome c[59]. Rat1, on the other hand, has been previously characterized as a protein that is found in the chloroplasts of *Chlamydomonas reinhardtii* which is involved in the splicing of group II introns found in the psaA mRNA that encodes a component of photosystem I. Rat1 binds an RNA cofactor named tscA and along with other protein factors mediates the maturation of psaA mRNA. Interestingly, the importance of the Rat1 protein in this mechanism is only apparent when a mutation is introduced in the tscA RNA cofactor. This mutation interferes with the psaA mRNA maturation and the correct formation of photosystem I unless the Rat1 protein along with other protein cofactors are present[61].

MCUR1 and Rat1 are connected for the first time in the present study, as they are both members of the mempromCC protein family. These two proteins are distant homologs and they share a well-conserved core in their head domain. Even though their previously described functions seem different in both cases they are involved in the correct formation of a protein that participates in the electron transport chain in either mitochondria or chloroplasts.

The interaction of Rat1 with Group II introns is of special interest. Group II introns are an ancient class of ribozymes that are found in bacteria, archaea as well as mitochondria and chloroplasts. They have the ability to cleave themselves, possibly with the assistance of proteins and they are considered to be the ancestors of nuclear gene introns[62]. Group II introns are characterized by their great sequence variation that makes them very difficult to be detected. This variation along with their localization in bacteria, archaea, mitochondria and chloroplasts are features that the Group II introns share with mempromCC-like proteins. These common characteristics might indicate an evolutionary and functional link between group II introns and mempromCC-like proteins.

The important but indirect involvement of Rat1 with the group II intron maturation mechanism which is only pronounced when there is a mutation in the tscA RNA cofactor, provides further support for a connection between group II introns and mempromCC-like proteins, since it could explain the great difficulty of detecting a clear phenotype for various mempromCC knock outs. A way to further support this hypothesis would be the verification of a cofactor similar to the tscA RNA for the other proteins found in the mempromCC family, something that was not possible until now.

Finally, the great variation in functionality, even within highly conserved mempromCC protein members, indicates how divergent evolution can alter the functionality of a protein. As mentioned in the paper "MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles" even the human paralogs MCUR1 and CCDC90B interact completely different with the MCU protein and with Ca^{2+}. MCUR1 has the ability to bind Ca^{2+} through the head domain, upon which it seems to structurally destabilize, while CCDC90B does not seem to interact with Ca^{2+} at all. The difference in functionality of three known members of the mempromCC family showcases mempromCC as a nice paradigm for divergent evolution.

Conclusions

MempromCC proteins belong to an evolutionary connected group of proteins that share the same unique secondary structure of an α-helical head, a β-layer neck, a coiled coil stalk and a transmembrane anchor found at the C-terminus of the proteins. MempromCC proteins are mostly found in proteobacteria and mitochondria of higher organisms and they share great sequence and structural similarities especially in the head-neck part of the proteins. The head is characterized by a highly conserved core of two α-helices separated by a conserved loop, a structure that is always found right before the β-layer neck. The function of the mempromCC proteins remain elusive since even the most closely related proteins show different functionalities as we describe in "MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles". On the other hand, the proteins studied like MCUR1 remain a subject of debate since they are involved in both the control of Ca^{2+} uptake and the electron transport chain in humans. One could assume that only one of these two hypothesis is correct, however, in both cases the MCUR1 protein seems to play an important role in the assembly of two proteins, the uniporter MCU and cytochrome c.

My hypothesis is that mempromCC proteins are universal assembling factors that are constantly altered by evolution to fit different needs. The detection of another more distant mempromCC member called Rat1 makes a compelling case for this theory. Rat1 is the only protein that was found in chloroplasts and even though it shares the same basic features with the rest of the family it is quite unique in function since it has been found to bind RNA and to mediate the alternative splicing of one of the components of photosystem I, another electron

chain transport protein. This mechanism of protein maturation involves group II introns and it allows for bypassing of mutations that would interfere with the correct assembly of photosystem I. We once more observe a strong connection of mempromCC proteins with the assembly of proteins critical for cell survival.

The interaction with group II introns is also of a great significance since they are the descendants of ancient riboenzymes.. The origin of mempromCC proteins seems to be very ancient and it presents a great example of how divergent evolution can drive the function of proteins in different directions. It is not lost on us that the other HNSA proteins could also be connected to mempromCC proteins via an ancient evolutionary path. However, large sequence and structure differences in the most conserved part of the proteins, the head-neck area, do not indicate an evolutionary connection. In this case, the formation of this characteristic HNSA secondary structure is a product of structural convergent evolution, highlighting the importance of not only this structure but also of the α/β fold formed by the head-neck. This new fold has been found in many other proteins as described in the paper "α/β coiled coils"[29].

Evolving a Novel Histidine Kinase

Project Summary

Histidine kinases are the most wide-spread sensory molecules of bacteria where they connect environmental signals to the expression of specific genes through systems known as two component signal transduction (TCST) systems. The catalytic-ATP binding domain CA that is present and highly conserved in both type I and II histidine kinases has a very interesting evolutionary story. The same domain can be detected fused with different domains in mitochondrial serine kinases[51,52], ATPases[53], DNA gyrases[54] and the DNA mismatch repair protein MutL[55], while it has also been found independent from other domains in a homodimer with a serine kinase function, called SpoIIAB[6,49,50].

In type I histidine kinases the CA domain is connected C-terminally to the dimerization and histidine phosphorylation (DHp) domain, forming the transmitter domain of the TCST system (Figure 4.A.). The DHp-CA transmitter is a homodimer found in the cytosol, which in the presence of a signal can autophosphorylate. During the autophosphorylation the γ-phosphate of the ATP bound to the CA domain is transferred on a histidine residue of the DHp domain. This leads to the phosphorylation of a response regulator (RR) on an aspartate residue triggering the expression of specific genes (Figure 2.A.). Even though histidine kinases also have a sensor domain and the HAMP (present in Histidine kinases, Adenylate cyclases, Methyl-accepting proteins and Phosphatases) domain connected N-terminally to the transmitter, sensor addition seems to be a later evolutionary step since sensor domains show great variation and the HAMP domains are very wide-spread among other proteins. In contrast, the DHp domain cannot be found in other proteins or independent from the CA domain, indicating that the fusion of these two domains happened early in their evolution.

In this project I have created three fusion proteins combining the DHp domain of the EnvZ histidine kinase[36,37] found in *E. coli* with another ATP binding domain called DX (Figure 4.). DX is an *in vitro* evolved protein created from a random library of 80 amino-acid long peptides, randomly mutated and selected on the ability to bind ATP[63]. DX evolution simulates the evolution of the first proteins that randomly bound different ligands, or each other. By combining this unique protein with the natural DHp domain I aim to address multiple evolutionary questions.

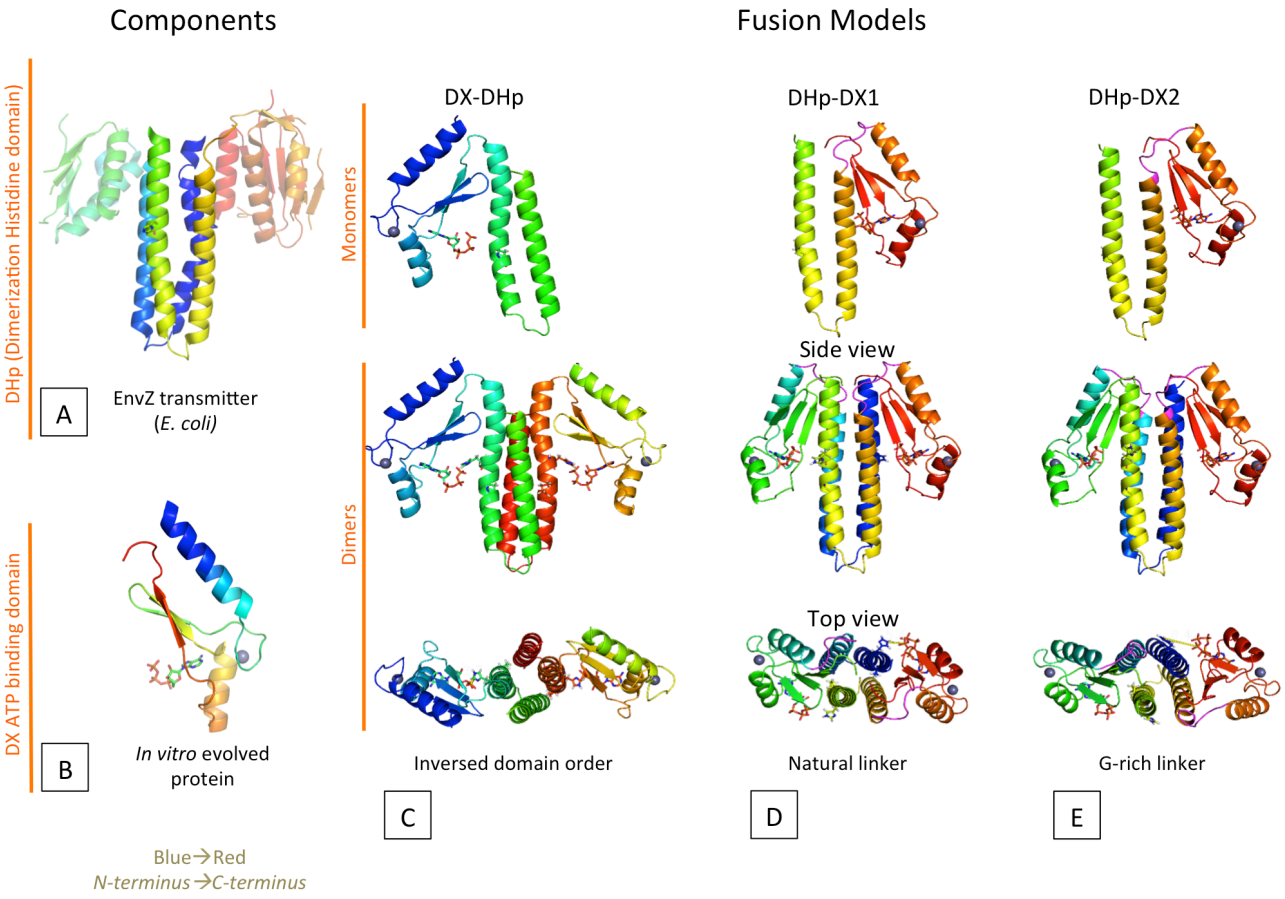


Figure 4. Design of fusion constructs. Three models combining the DHp and the DX domains. **(A)** DHp-CA transmitter highlighting the DHp domain. **(B)** *In vitro* evolved ATP binding DX domain. **(C)** DX-DHp protein with an inversed domain order. **(D)** DHp-DX1 protein with the EnvZ natural linker. **(E)** DHp-DX2 protein with a G-rich designed linker. The monomeric and dimeric forms (front and top view) of the fusion proteins are shown from top to bottom. The rainbow color from blue to red indicates the N to C-terminal orientation of the proteins. The loops are shown in magenta.

Firstly, we simulate the creation of a primordial enzyme that could have evolved by the combination of different domains leading to an low efficiency though evolutionary important enzymatic functions. Secondly, we recreate the evolution of type I histidine kinases by using a simple, randomly evolved ATP binding domain in place of the CA domain. This combination simulates the first fusion of CA and DHp domains, creating a precursor of histidine kinases. We address these questions in the enclosed paper "Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase.", where we show that one of our DHp-DX fusions showed a well-characterized and specific ATPase activity similar to natural ATPases. The dephosphorylation of ATP to ADP is part of the function of histidine kinases, since it is the first step to phosphorylate the histidine on the DHp domain. As expected from a precursor of a histidine kinase, part of the final function can be found in our fusion protein DHp-DX1, even though the phosphogroup does not seem to be transferred within the protein or on the response regulator OmpR, *in vitro.*

The function of the DHp-DX1 protein i*n vivo*, though, seems different. I have established a modified TCST system that combines the Taz and Cph1/EnvZ chimeras (Figure 2B, C) allowing the coupling of all three fusion proteins to GFP expression. This system showed that only the DHp-DX1 fusion protein led to a strong gene expression up-regulation through the ompC promoter, indicating that the protein functions as an OmpR phosphorylase in the cell. This hypothesized phosphorylase activity is further supported by the finding that the observed gene expression up-regulation via the ompC promoter is completely reversed for the histidine deficient mutant DHp-DX1(H15Q).

The lack of *in vitro* evidence poses questions regarding the kinase function of the DHp-DX1 fusion protein, however an established *in vivo* system that facilitates the readout of the OmpR phosphorylation is highly important since it allows the construction of an *in vivo* evolution mechanism for the selection of mutations which benefit the kinase functionality. I managed to couple the DHp-DX1 system to a reporter gene for ampicillin resistance that allows the selection of any increase in the phosphorylase activity under semi-lethal conditions. Random

mutagenesis of the DHp-DX1 fusion protein is now attempted in combination with the constructed *E. coli* strain JM_Amp_LVA, aiming to improve any kinase activity of the protein. This final attempt could show that an ATPase was a precursor to histidine kinases, while in parallel it could show that the selection of the CA domain was a random, opportunistic event, even if it later was selected in multiple proteins with widespread functionality. A finding like that offers a proof of concept for convergent evolution, where two truly analogous proteins could replace each other into a functional system. Finally, this project introduces the concept of combining *de novo*-*in vitro* evolution, natural domain utilization or "domain recycling" and *in vivo* selection as a method to create novel enzymes, aiding the process of enzyme design.

Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a *de novo* evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase.

Karamichali I, Sepulveda E, Zhu H, Coles M., Lupas AN.

(Ready for submission; see Appendix)

Abstract

We recreate the evolution process of a primordial enzyme by adding the natural dimerization and histidine phosphorylation (DHp) domain to an *in vitro* evolved ATP binding protein (DX), creating a fusion protein that simulates the cytosolic EnvZ histidine kinase of *E. coli*. The fusion protein successfully binds and hydrolyzes ATP *in vitro*, while significantly affecting the ompC promoter mediated gene expression *in vivo*. This is the first time a *de-novo* evolved protein has replaced a natural domain within the context of a whole functional protein system *in vivo*. This finding provides a proof of concept of the evolution of primordial enzymes, as well as the evolution of histidine kinases. Our approach can be further applied for the creation of novel enzymes and the facilitation of novel enzyme design.

Contribution

I was the main contributor of this paper for which I produced many of the ideas, designed and executed most of the experiments and analyzed the data. I also prepared the manuscript and the figures, which were critically examined and edited by Dr. Edgardo Sepulveda and Prof. Dr. Andrei N. Lupas.

The other authors contributed as indicated here:

1. Edgardo Sepulveda: Input and ideas regarding the *in vivo* experiments
2. Hongbo Zhu: Input and ideas regarding the *in silico* protein design
3. Murray Coles: NMR experiments
4. Andrei N. Lupas: Conception of the project

Additional Methods

Increasing protein stability

All the proteins used were fused N-terminally with thioredoxin (TRX) to increase their stability *in vitro*. This was accomplished by cloning the genes in the pETTRX vector, which includes a 6xHis-tag placed between TRX and the protein of interest. The restriction sites NcoI and BamHI were used for all the fusions. Fusions with only the 6xHis-tag or the $HAMP_{Af1503}$ domain (HAMP domain found in a sensor-HAMP only system of *Archaeoglobus fulgidus*[46]) known for its ability to stabilize proteins, were also tested resulting in insufficient soluble protein expression. The fusion proteins DHp-DX1, DHp-DX2 and the mutant DHp-DX1(H15Q) were stabilized during overexpression by inducing with 2% ethanol. Buffer screening established buffer A (30mM Tris, 200mM NaCl, 5mM citric acid, 5mM 2-mercaptoethanol, 5μM $ZnCl_2$, 10% glycerol, pH 8.5) as best for further stabilization of the fusion proteins, even during and after the cleavage of the TRX-6xHis fusion with TEV protease.

The establishment of the overexpression conditions of the DX protein required an extensive screening. 24 of the 72 selected expression conditions were proposed by the online tool SAmBA[64]. This tool offers the minimal number of conditions needed to equally represent each variable of interest. The cells were grown and induced with 0.25mM IPTG ± 2% ethanol or 0.2mM NaCl within a 24well plate either at 30 or 37°C. Multiple cell lines (*E. coli BL21gold, E. coli C41 and Rosetta 2 (DE3) pLysS*) and fusion constructs were tested The fusion constructs were prepared by cloning in the plasmids pET-15b, pET28a-C(+), pETHIS, pRTTRX for adding no tag, a C-terminal 6xHis-Tag, a N-terminal 6xHis-

Tag and finally a N-terminal TRX-6xHis-Tag, respectively. After 4 hours the plates were centrifuged for 10min at 4000rpm and frozen in -80°C after the removal of the supernatant. I lysed the cells by thawing and resuspending in lysis buffer [Buffer B: 30mM Tris, 500mM NaCl, pH 7-8.5 complemented with Lysozyme (Sigma Aldrich), DNase I (AppliChem), and 1 Protease Inhibitor Tablet (cOmplete Roche)], within which they were incubated for 1 hour at RT and under shaking. Samples were centrifuged at 4000 RPM for 8min and a 10% SDS gel was used to check the expression levels (10μl of sample plus 10μl of 2x SDS loading buffer: 62.5mM Tris, 2% SDS, 10% glycerol, 5% 2-mercaptoethanol, 0.001% bromphenolblau, pH equal to 6.8). The best overexpression condition was selected based on the intensity of the bands in the soluble fraction. The best results were obtained when Rosetta 2 (DE3) pLysS cells were used during the induction of 0.2mM NaCl at 37°C. The proteins DHp-CA, and the mutant DHp-CA(T19Q), which is known to have a much lower phosphatase activity[65], were overexpressed as described in table 2 and 3 (see Appendix). The stabilization of the proteins during freezing was accomplished by addition of either 50 or 5% glycerol, depending on the application for which they were stored.

Stability assessment

Any increase in protein stability was assessed using size exclusion chromatography and circular dichroism (CD). Buffer A or C (30mM Tris, 150mM NaCl, pH 8.5) was used during size exclusion chromatography of purified protein samples followed by a nickel affinity chromatography and a dialysis step. The CD analysis was performed with 0.3mg/ml protein dialyzed in 30mM Tris, 200mM NaCl and pH 8.5. The spectrum was measured with the J-810 Spectropolarimeter JASCO from 201-250nm. A melting curve was also performed at 220nm between 10-95°C.

Crystallography

The crystallization of the protein DHp-DX1 with and without the TRX fusion was extensively attempted (96 different conditions) using the sitting drop method in the presence and absence of ATP, $MgCl_2$, and $ZnCl_2$.

NMR

Cleaved from TRX and re-purified DHp-DX1 protein (7mg/ml) was analyzed with one-dimensional NMR spectroscopy in Buffer C. Further analysis was not possible due to protein stability problems caused when M9 minimal media (+ Traces, 1µg/ml biotine and thiamine) was used during the ^{15}N labeling of the protein.

Mass Spectrometry

Mass spectrometry (MS) was performed as part of the *in vitro* functional characterization of DHp-DX1. TRX-fused DHp-CA protein samples were tested at the Proteome center in collaboration with Prof. Dr. Macek, to check whether histidine phosphorylation can survive the acidic conditions of the MS protein preparation. The TRX-fused DHp-CA protein samples were prepared after Ni^{2+} column purification. The protein was tested both immobilized in a blue native gel (in the presence and absence of 4mM ATP) and in solution (buffer B). No histidine phosphorylation was detected even if the phosphorylation was detectable by western blot (see paper "Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase."). No further analysis was performed.

Autoradiography

The TRX cleaved proteins DHp-CA(T19Q) and DHp-DX1 were stored in buffer D (0.1mM Tris, 50mM KCl, 5mM β-mercaptoethanol, 5% glycerol) and were mixed

with the OmpR response regulator in concentrations between 2-20μM. The samples were screened under the addition of either $MgCl_2$ and/or $CaCl_2$ or $MnCl_2$ (5-10mM) in the presence of a mixture of cold (5-50μM) and radiolabeled (0.2-1μl of gamma-^{32}P ATP; 10mCi/ml). Buffer D was used adjusted to pH of 8, in order to mix the selected metals with either the DHp-CA(T19Q) or the DHp-DX1 protein. The incubation started with the introduction of the cold/hot ATP mixture and lasted for 15 min at RT. The OmpR response regulator was added for an additional incubation of 0, 1, 5, 15, 30 and 60 minutes. A sample containing only one of the proteins (OmpR) was used as a negative control. The reaction was stopped by adding 4x loading buffer and by boiling the samples for 2 min. The samples were finally loaded on either an in-house (18% polyacrylamide, in Anode/Kathode buffer: 25mM Tris, 0.2M glycin, pH 8.5) or commercial SDS gel (12% Invitrogen, in running buffer: 50 mM MOPS, 50 mM Tris Base, 0.1% SDS, 1 mM EDTA, pH 7.7). The gel was run at 25-30mA for about 1.5 hours. The gel was washed, sealed and analyzed via autoradiography (24h exposure time, visualized by the Fuji Imager FLA 3000).

Isoelectric focusing (IEF)

The ability of the fusion protein DHp-DX1 to bind OmpR compared to the positive control DHp-CA was tested similarly to the autoradiography protocol, but only using cold ATP. The gels used were commercial isoelectric native gels (pH 3-10 IEF Gel; Invitrogen), The samples were mixed with loading buffer (20mM free base Lysine, 20mM free base Arginine, 15% Glycerol) and loaded (15μl) on the gel, alone or in combination with OmpR. An anode (7mM phosphoric acid) and a cathode buffer (20mM free base Lysine, 20mM free base Arginine, pH 10.1) were used to run the gel at 100V (7-4mA) for 1 hour. The effect of the metals $MgCl_2$, $CaCl_2$ or $MnCl_2$ (10mM) on the binding was also tested. The same experiment was performed using also native gels (4-16% NativePAGE ™ Bis-Tris; Invitrogen).

In vivo evolution strain construction

Different *E. coli* strains were constructed based on the JM1012 strain provided by Lee *et al.* (2013; Figure 2.C.). The repressor gene Cl_LVA that was controlled by the ompC promoter was replaced with either a kanamycin or an ampicillin resistance gene with or without the LVA protease recognised-11-amino-acid long tail (Figure 10.A. and B.). The strains constructed were JM_Kan_LVA, JM_Amp_LVA and JM_Amp.

In parallel the toxic, in the presence of sucrose gene SacB was also used to replace the GFP gene in the pJM1 and 2 plasmids, also provided by Lee *et. al* (Figure 10C.). The plasmids constructed were pJM1_SacB, pJM1_SacB_LVA, pJM2_SacB and pJM2_SacB_LVA.

All constructs were made by PCR and a lambda red recombinase (λRed) mediated homologous recombination directly into the JM1012 cells. The primers for each PCR were designed with a 50 nucleic acid-long overhang that was homologous to either the 5' or 3' end around the desired position of cloning. The kanamycin and ampicillin resistance genes were cloned into the chromosome in place of the Cl-LVA gene downstream of the ompC promoter. Similarly, the SacB gene was cloned in either of the plasmids pJM1 and 2, replacing the GFP gene downstream of the P_L promoter. The PCR products were purified and cloned into electrocompetent cells through electroporation at 1.7V.

Electrocompetent cells were prepared using JM1012 cells cloned with the λRed plasmid pCP20 (thermosensitive over 30^{o}C) and if needed either the pJM1 or pJM2 plasmid (only for the SacB transformations). The transformed cells were grown in 30^{o}C until at OD of 0.3, and then they were induced with 10mM arabinose in order to express the λRed recombinase. After 4h the cells were washed with ice cold water and then glycerol was added to reach 10% (w/v). The cell preparation was finally aliquoted (50μl) and frozen in liquid nitrogen for storing at -80^{o}C.

The successful transformation of the cells was verified through colony PCR using the same primers used for the preparation of the genes. The colonies were selected after the induction of the ompC promoter using 2.5%(w/v) NaCl that rendered the cells resistant to either kanamycin or ampicillin, depending on the

gene[40]. The selection of the SacB transformants was based on the color of the colonies since the gene replaced was the GFP gene leading to the formation of GFP negative clones.

Strain selection characterization

All strains described above were further tested under different concentrations of either kanamycin, ampicillin (Roth) or sucrose (Sigma) to determine their effectiveness in selecting any OmpR kinase functionality. The strains were tested in the absence or presence of one of the proteins Taz, DHp-DX1 and DHp-DX1(H15Q). All the proteins were cloned into the electrocompetent cells of each strain (see above) carried in pCL1920 plasmids.

The samples for screening were prepared by inoculating 5ml LB with cells from a glycerol stock. 50μl of the overnight cultures (30°C) were transferred into 5ml LB, which was incubated under 37°C for 2 hours. The OD of the culture was adjusted to 0.3 (570nm) and 200μl were placed in each well of a 96 well plate (with clear bottom, from Sigma-Aldrich). The plates were incubated at 37°C overnight, while the OD was monitored every 15min at 570nm. The screening was performed in the plate reader Synergy H4 or MX Hybrid Reader (BioTek Instruments). The screening that included bacteria with the pCL1920 plasmid were performed in the presence of spectinomycin (Sigma).

DHp-DX1 random mutagenesis

The random mutagenesis of the DHp-DX1 protein was accomplished by error prone PCR as described by McCullum *et al.* (2010)[66]. The primers used where:

pCL1920_flaging_DHp-DX1_F
ATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCTTG
pCL1920_flaging_ DHp-DX1_R
GTTGTAAAACGACGGCCAGTGAATTCGAGCTCGGTACCCGGGGATCCTTA

which have 50 nucleic acid long 5' and 3' homologous to the pCL1920 plasmid overhangs. As described above these overhangs allow the direct transformation into the plasmid through the λRed homologous recombination, replacing the original DHp-DX1 gene with any of the mutants. The DNA used was bacterial extracted pCL1920(DHp-DX1) plasmid (50ng/μl).

In vivo evolution

Competent cells of the stain JM_Amp_LVA transformed with pCL1920(DHp-DX1) and the λRed pCP20 plasmid were prepared as described above. Chemocompetent cells were also prepared by adding 0.1M $CaCl_2$ in the cell preparation and incubating for 30min on ice before aliquoting. The purified random mutagenic PCR products were finally used to transform the cells. The following incubation of the cells in 30^{o}C for 2 hours allows the λRed mechanism to directly clone the PCR products into the chromosome in place or the original DHp-DX1 gene. The selection was based on semi-lethal conditions (50μM) of ampicillin in freshly prepared plates that also included spectinomycin.

Additional Results/Discussion

Increased protein folding during overexpression

The purification of all the DHp and DX fusion constructs was challenging because of their low solubility. The TRX fusion greatly improved the stability of the DHp-DX1 and DHp-DX2 proteins, while it was not efficient in stabilizing DX-DHp. The buffer used was important for keeping the proteins DHp-DX1 and DHp-DX2 from precipitating during the concentration or the TEV protease cleavage of the TRX fusion.

The $HAMP_{Af1503}$[46] domain from *Archaeoglobus fulgidus,* that is known for its stability was also used fused to the N-terminus of the DX-DHp, DHp-DX1 and DHp-DX2 proteins to stabilize them and facilitate their purification. These constructs showed some overexpression in Arctic and C41 cells, however, the solubility was not improved.

Further testing showed that the induction with 2% ethanol during the overexpression of the TRX fused DHp-DX1, lead to expression of a highly stable protein. The addition of ethanol stressed the cells, leading to the expression of the heat shock proteins GroES/EL and DnaK/J that bind unfolded proteins, facilitating their correct folding[67]. Further analysis with size exclusion chromatography and circular dichroism (CD) showed that the protein was well folded, it was probably forming a dimer and had a melting point at 67^{o}C (Figure 5 and 6).

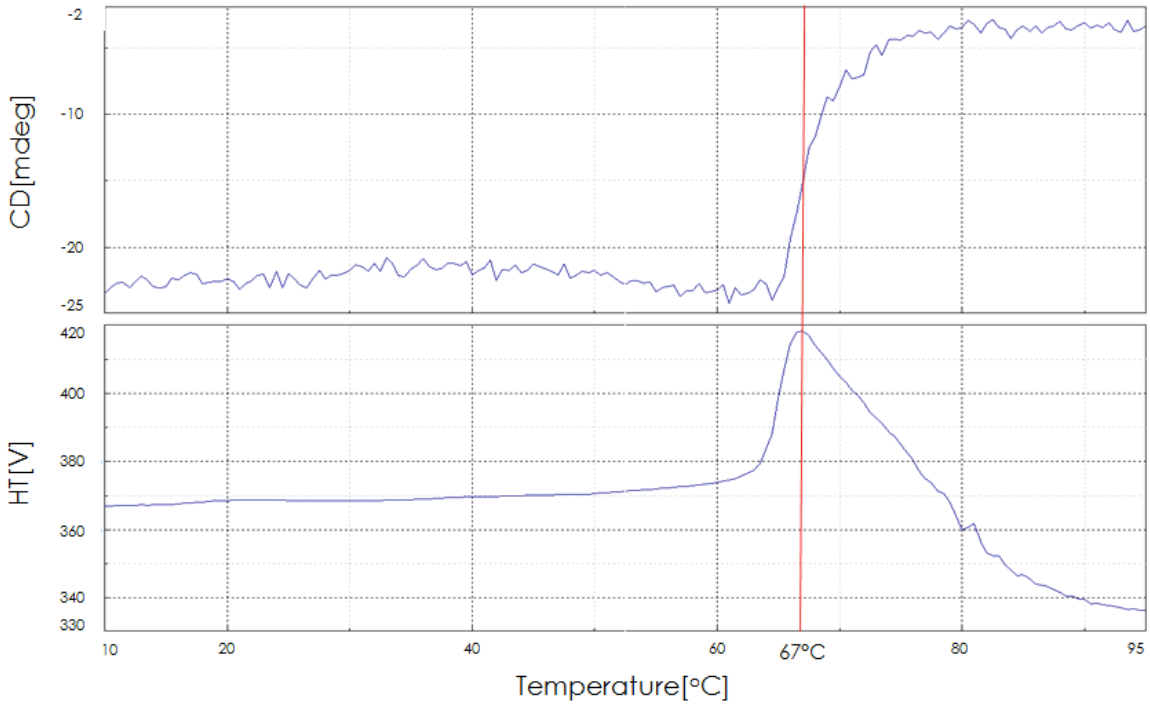


Figure 5. CD melting curve of DHp-DX1 fused with TRX. The protein is well folded and thermostable up to 67°C after expressing in the presence of 2% ethanol.

The huge improvement in protein stability is illustrated by the results of the size exclusion chromatography as shown in figure 6. Before the use of ethanol (Figure 6.A.) most of the protein was eluted in the void volume as inclusion bodies. In contrast, the use of ethanol led to a well-folded protein that eluted predominantly at the expected size of 60kDa (Figure 6.B.). This observation also suggests formation of a homodimer of the fused to TRX DHp-DX1 protein. The same expression method was used for the purification of the TRX fused DHp-DX1(H15Q) mutant, with the same result. The expression of soluble DHp-DX2 protein was also improved using the same approach, however the effect was not as prominent.

The positive controls DHp-CA and the mutant DHp-CA(T19Q) were soluble well-folded proteins that caused no problems during any step of the expression, purification or storage. The DX and OmpR proteins, on the other hand, were challenging. OmpR has been studied many times in the past and it is known that the fusion with a TRX protein in combination with a very slow expression is ideal

for this protein[68]. Nevertheless the lifetime of the protein was quite short since it was precipitating at a fast rate even on ice.

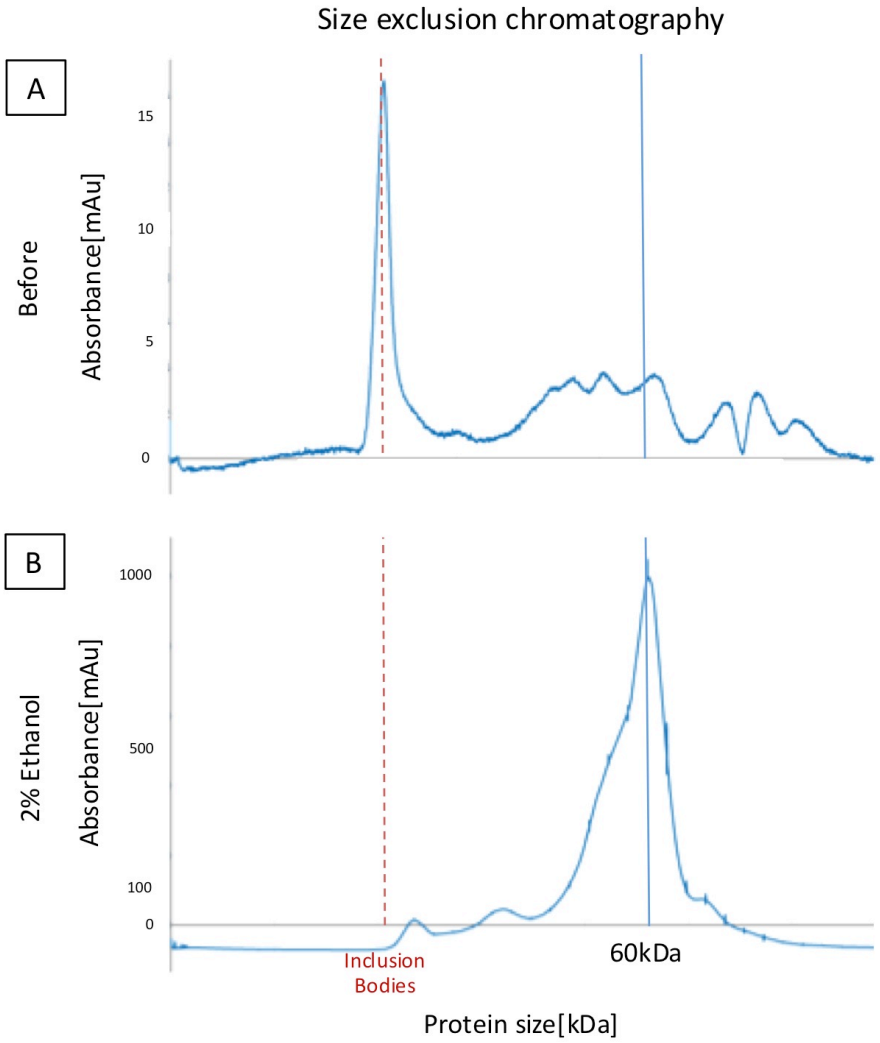


Figure 6. Size exclusion chromatography. Results before **(A)** and after **(B)** using 2% ethanol during overexpression.

The overexpression of DX was extensively tested using a statistical approach, which allowed the evaluation of 72 different conditions, including 3 different cell lines (BL21 Gold, C41, Rosetta 2 (DE) lysS), 3 fusion states (±C-teminal or N-terminal His-tag, or N-terminal TRX fusion with a His-Tag), 3 additive conditions (±2% Ethanol, 0.3M NaCl) and two temperatures (30,37°C). While a test of all the possible combinations (full factorial analysis) would be expensive and time consuming, the statistical approach allowed a fast and economic evaluation of all

possible variables by testing only 24 different conditions[69,70]. The representative conditions were selected using the online tool SAmBA[64], which is a software that optimizes the design of biological macromolecule experiments, originally designed for the optimization of crystallography condition [64]. The 24 conditions were selected to equally represent all different variables, allowing the fast-in lab testing of the significance of each variable. The DX protein was best expressed fused with TRX in the Rosetta cell line (see table 2,3 in the Appendix) under 37^{o}C. The identification of these conditions lasted a week, given the plasmids for cloning were ready for use.

Structural analysis

Both crystallography and NMR were employed to structurally analyze the DHp-DX1 protein. The crystallization of DHp-DX1 with and without the TRX fusion or the ethanol induction was extensively attempted, but no protein crystals were formed. The plates were monitored continuously over time. A large number of crystals were observed after two years, some of which are currently under analysis.

The DHp-DX1 protein cleaved from the TRX fusion as described above was also used for NMR analysis. The unlabeled protein was folded and found to form a dimer. The ^{15}N labeling of the protein was also attempted, however the expression of the protein within minimal media provoked folding problems even in the presence of 2% ethanol.

In vitro functional analysis

In addition to other techniques described in the paper "Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase." the DHp-DX1 was also tested in its ability to bind the response regulator OmpR, in comparison to the positive control DHp-CA. The test was attempted in both regular native gels and in native-isoelectric focusing (IEF) gels. The result was similar in both cases with IEF gels providing a much

clearer picture. In the case of the positive control a new band was observed when the DHp-CA protein was mixed with OmpR (Figure 7). This band was clearly different from the bands of the single proteins and it is the band of the complex forming between the DHp-CA protein and the response regulator OmpR. An analoge to this band was not observed when DHp-DX1 was tested. Different metals ($MgCl_2$, $MnCl_2$ or $CaCl_2$) were tested in both cases but showed no effect on the formation of a complex.

This observation means that the fusion protein DHp-DX1 has a lower affinity to the response regulator in comparison to the positive control even if the DHp domain, which is the domain responsible for the interaction, remained the same. This might be one of the reasons for the low efficiency of the protein to phosphorylate OmpR as observed during the autoradiography experiments.

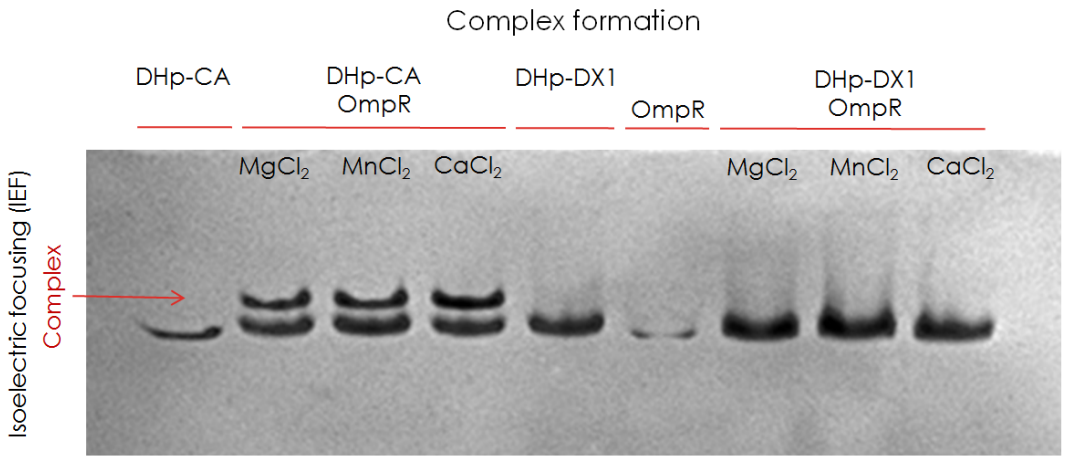


Figure 7. Isoelectric focusing native gel. The fused protein DHp-DX1 could not bind the response regulator OmpR as efficiently as the positive control DHp-CA. The addition of 10mM $MgCl_2$, $MnCl_2$ or $CaCl_2$ had no effect in the formation of the complex.

Mass Spectrometry

Histidine and aspartate phosphorylation, differ from the phosphorylation of serine, threonine and tyrosine since they create a phosphoramidate (histidine) or an acyl-phosphate (aspartate) instead of a phosphomonoester (serine, threonine, tyrosine). Phosphomonoesters are stable under acidic conditions, while phosphoramidates and acyl-phosphates are acid-labile and regularly

missed during analytic methods that demand acidic conditions[71]. Mass spectrometry (MS) demands such conditions, even though there have been examples were histidine phosphorylation was successfully determined[71–73]. MS is also important since it can detect the exact phosphorylation site either in a phosphorylated or a dephosphorylated state, a feature that could allow the determination of the auto-phosphorylation turnover of the kinase. To test whether this method could work for our fusion constructs, we tested the technique first on the TRX fused positive control DHp-CA, which has been proved to have phosphorylated histidine via western blot (enclosed paper "Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase."). Unfortunately, no histidine phosphorylation on the positive control could be detected and therefore no further experiments were performed.

Autoradiography

Autoradiography is one of the most sensitive techniques for the detection of any phosphorylation. We have extensively attempted to visualize any phosphorylation on the fusion protein DHp-DX1 or the OmpR response regulator by using gamma-^{32}P ATP at different protein concentrations, ratios of cold/hot ATP, or metals ($MgCl_2$ and/or $CaCl_2$, or $MnCl_2$). During these attempts we encountered many problems affecting the success of the tests. The main problem was that we had to screen many conditions, which might have worked for the fusion protein DHp-DX1 but they are not suitable for the positive control, rendering us blind during many of our experiments.

DHp-CA and OmpR have been well studied through the gamma-^{32}P ATP labeling[65,74,75], however usually the ATP was removed from the sample after the autophoshorylation of the kinase and before the introduction of the response regulator OmpR. This avoids background noise and increases the sensitivity of the analysis. This approach also allowed to distinguish between the autophosphorylation of the DHp-CA kinase and the phosphorylation of the OmpR

response regulator, two proteins which are very close in size (24.7 and 28kDA, respectively) making the acrylamide gel separation very difficult. In our case, the low efficiency of the fusion protein DHp-DX1 and the probably different mechanism of OmpR phosphorylation (enclosed paper "Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase."), forced us to mix the ATP and both proteins together. This led to an increase in background and also a constantly phosphorylated DHp-CA control that was making the signal separation from the one of OmpR very hard (Figure 8). The TRX fused proteins were also used to avoid this problem, however it was obvious that even if the autophosphorylation of the positive control was not affected, the phosphotransfer onto OmpR was severely affected even if the TRX fusion was only present on the kinase. In parallel, it has been well established that the phosphatase activity of DHp-CA can rapidly dephosphorylate OmpR making the detection of the phosphorylation impossible unless $CaCl_2$ is used in place of the $MgCl_2$. To be able to test all metals we also constructed the DHp-CA(T19Q) mutant which is known for the much lower phosphatase activity[65] in comparison with the native DHp-CA protein.

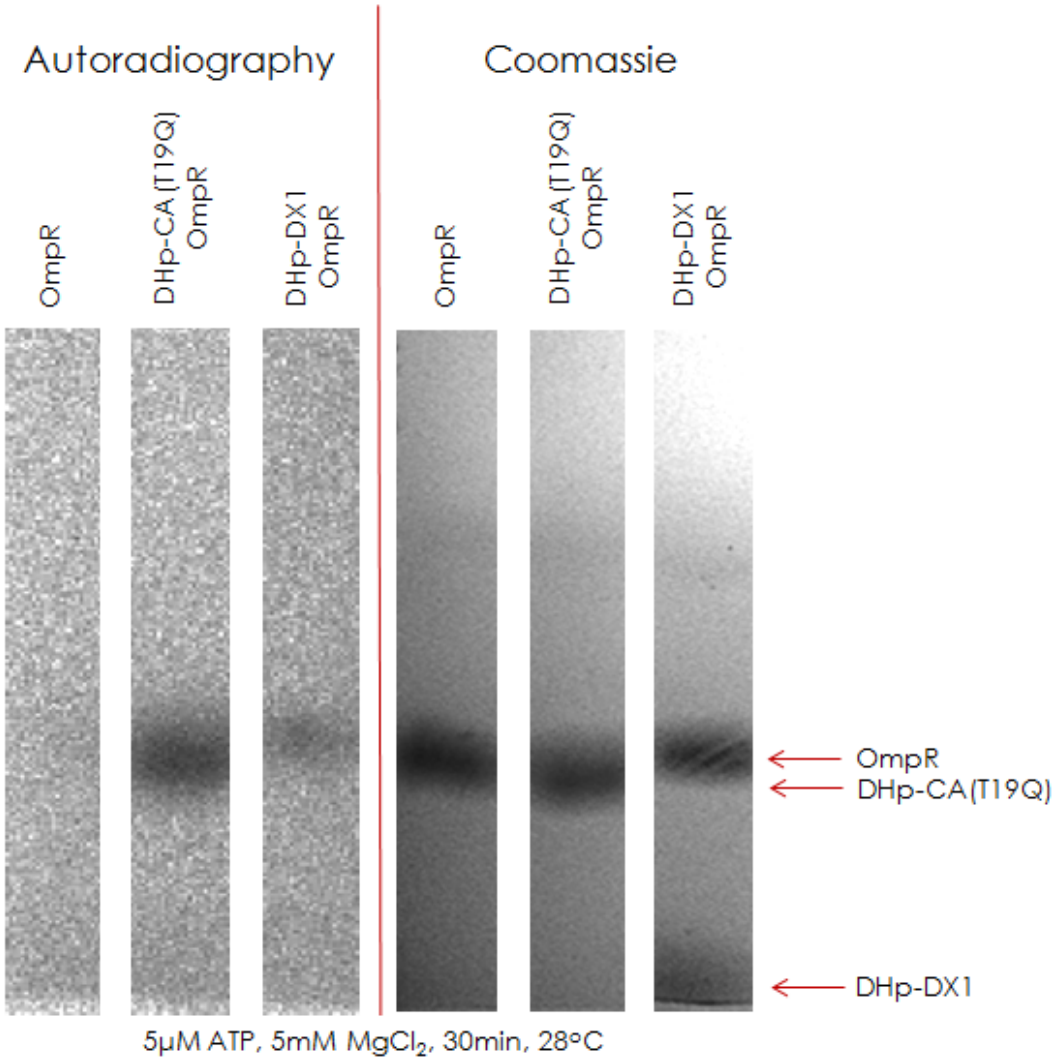


Figure 8. Autoradiography results. Radioactivity based Imaging after 24h of exposure (left). The gel lines of the samples with only the OmpR protein, a mixture of OmpR and the positive control DHp-CA(T19Q) and finally the mixture of OmpR with the fusion protein DHp-DX1 are presented from left to right. The same gel was placed in coomassie after imaging (right). The signal of DHp-CA(T19Q) interferes with the one of OmpR. No autophosphorylation is detected for the fusion protein DHp-DX1, while there is clear but low phosphorylation of the OmpR response regulator in the same sample. The conditions of the experiment were 5μM ATP, 5mM $MgCl_2$, 28°C and incubation time 30min. The experiment could not be reproduced.

After many attempts we were not able to detect any autophosphorylation of the fusion protein DHp-DX1, nevertheless we could detect a weak but very clear phosphorylation of OmpR when incubated with only the fusion protein (Figure 8). This observation could mean that we had indeed successfully produced a novel histidine kinase which has a short histidine phosphate half-life, making it hard to detect. Alternatively, this could mean that the protein acts as an

aspartate kinase that directly phosphorylates the OmpR response regulator. There are ways to investigate these theories further. Currently, we unfortunately could not reproduce this result. There are many reasons why this could be the case and further screening could establish in greater details the exact conditions needed to get a positive result.

In vivo functional analysis

In this project the chimeric Taz TCST system (Figure 2.B.) constitutes the backbone of the *in vivo* functional analysis. The system was preliminarily tested with both the gfp_LVA and the wild type GFP reporter genes. However, the system was characterized by high background noise, which had been also observed in the past, and it was affected by different osmolarity conditions and different temperatures used in the pre-induction overnight cell culture (data not shown). A reason for the temperature dependence might have been the alteration of the copy number of the vectors used to carry the reporter system. Another source of the high background is the OmpR response regulator itself. First, the response regulator OmpR can be cross phosphorylated from other kinases, like the chemotaxis sensor histidine kinase CheA[76,77]. Second, OmpR can trigger gene expression through two different promoters, the ompC and ompF promoter, depending on the osmolarity applied to the system[76]. High osmolarity triggers the gene expression through the ompC promoter, while a lower osmolarity redirects the signal towards the ompF promoter. Both promoters lead to the expression of porin proteins in nature, however only ompC is used for the expression of the reporter gene in the chimeric Taz TCST system, meaning that a low osmolarity will reduce the final signal of the system. In parallel, it has been reported that the EnvZ DHp domain might be capable of sensing osmolarity changes in the cytoplasm, altering the level of phosphorylated OmpR, even if the DHp-CA protein mostly acts as an OmpR phosphatase[78].

The complicated nature of this TCST system made it difficult to get clear and reproducible answers about the functionality of the DHp-DX fusion proteins, or even the controls. All the above made clear that it was of vital importance to

repeat the experiments in another, more tightly controlled cell system. Lee et al. (2013) introduced a system like that, in their effort to design a light-switchable TCST system which combines the photosensor Cph1 with the cytoplasmic part of EnvZ (Cph1-EnvZ, Figure 2.C.).

The Cph1-EnvZ chimera generally follows the same principles as the Taz TCST system, the sensor reacts to a signal, leading to either the phosphorylation or the dephosphorylation of the OmpR response regulator. The phosphorylated OmpR can then trigger the gene expression through the ompC promoter. The main difference between this new system and the Taz TCST system used before, is the tighter control of the reporter gene. In the first case, the ompC promoter is found in the chromosome instead of the plasmid, reducing the high background of the system. The background is further reduced by the dual control of the reporter system through the use of a highly controlled second promoter (P_L) repressed by the CI-LVA protein that is expressed under the ompC promoter. Finally, the reduction of the fluorescence signal as a result of the phosphorylation of OmpR, eliminates the noise created by any crosstalk phosphorylation of the same protein.

I have altered the Cph1-EnvZ chimeric system in order to fit the needs of our study (Figure 9.A.; enclosed paper "Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase."). While both strong and weak ribosome binding site (RBS) pJM vectors were tested, only the weak promoter gave us readable results. All fusion proteins were used in the new cell system under stable osmolarity conditions. Interestingly, this time the distinction between the phosphatase and the phosphorylase activity of the constructs was clear. By extracting the signal of the negative control, which in this case was again the same system containing no construct of interest, proteins with a phosphatase activity gave a signal higher than one, while the ones with a phosphorylase activity gave a signal below one. I selected to extract the background by using it as the base of the logarithm because the signal was exponentially altered depending on the background. Any other algorithmic representation does not clearly separate the positive and negative signal.

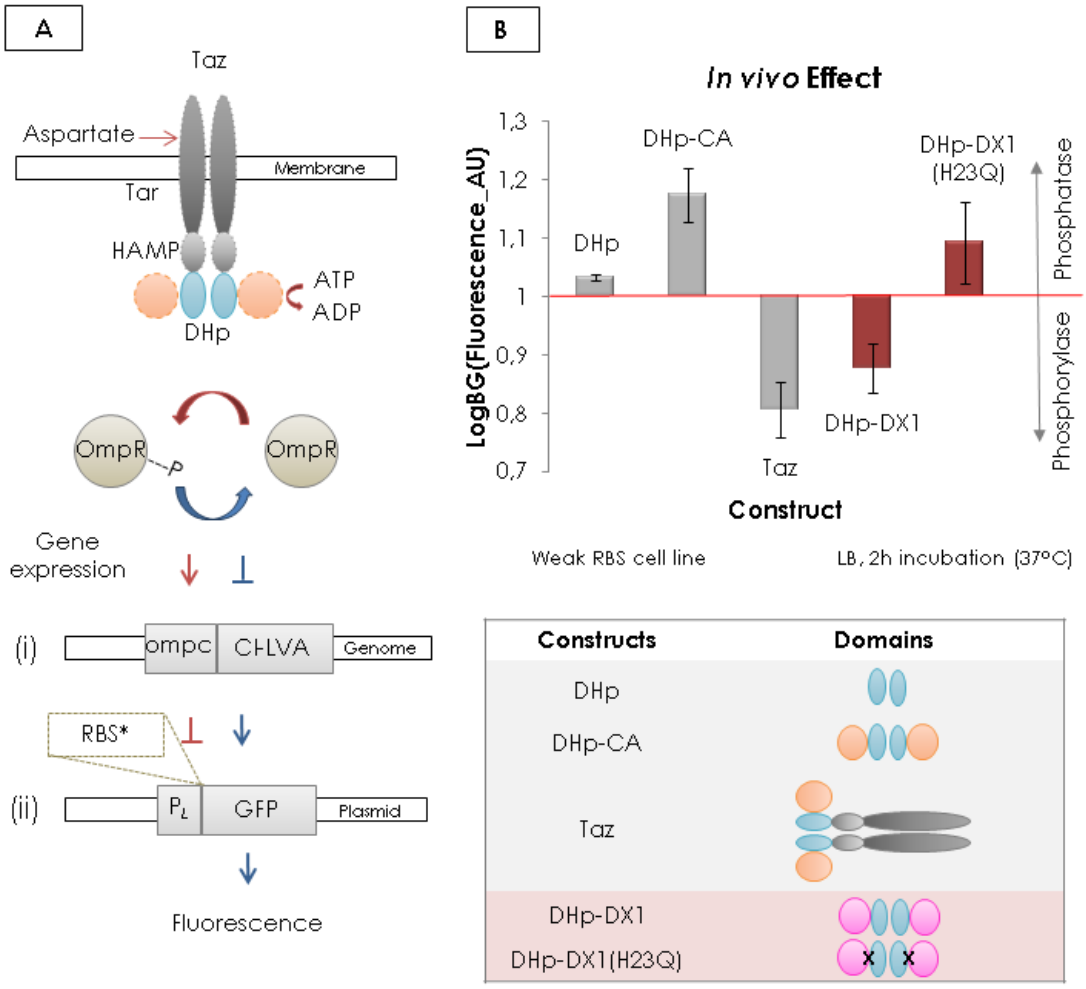


Figure 9. ***In vivo*** **functional analysis of DHp-DX1 fusion. (A)** Adapted Taz and CPh1-chimeric TCST systems, allowing the detection of both OmpR phosphorylation and dephosphorylation based on the dual control of the expression of the GFP reporter gene. In the first step the ompC promoter controls the expression of the CI_LVA repressor (i) which then regulates the always activated GFP expression by interacting with the P_L promoter (ii). A weak or a strong RBS could be selected (*). Domain removals or exchanges are shown with dotted outlines. **(B)** Graphic representation of the logarithm of GFP fluorescence, using the fluorescence of the background as the base of the logarithm. The assay conditions are shown at the bottom of the graph. phosphatase activity increases fluorescence (upward bars) while phosphorylase activity decreases fluorescence (downward bars). The constructs shown are listed in the table; grey represents controls while red indicates the fusion DHp-DX1 and the analogous histidine deficient mutant. DHp-DH1 showed histidine dependent phosphorylase activity *in vivo.* The experiment was repeated six times

Between all controls (shown in gray in figure 9.B.), Taz is the only one with a phosphorylation activity. Interestingly, any alteration of the system (removal of the sensor-HAMP domains, or removal of the ATP binding domain) switched the functionality of the protein towards a phosphatase activity, even when only the DHp domain was tested.

Surprisingly, when more constructs were analyzed, the proteins DHp-DX1 and DHp-DX2 showed also a phosphorylase activity. For DHp-DX2 the effect was really weak and seemed insignificant, while for DHp-DX1 the phosphorylase activity was almost half of the one observed for Taz. An interesting observation is that the addition of the Tar sensor to the DHp-DX1 protein through the HAMP domain (T-DHp-DX1), switched the protein towards the phosphatase functionality. This is the opposite effect from the one observed for the Taz and the DHp-CA protein.

This observation indicates that there is an effect of the DX domain on the functionality of the DHp-DX1 protein. Any alteration of the Taz system led to phosphatase activity, a characteristic of the DHp domain that is only reversed in combination with the Tar sensor. The only exception to this finding seems to be the DHp-DX1 protein. Furthermore, the histidine deficient DHp-DX1(H15Q) protein lacks any phosphorylase activity.

These findings strongly indicate that we have created a protein that *in vivo* controls gene expression through the ompC promoter in a similar way to Taz but with half of the efficiency. Moreover, this function completely depends on the histidine residue that usually gets autophorphorylated on the DHp domain. Thus, DHp-DX1 is a protein that probably predominantly phosphorylates OmpR while the analogous native DHp-CA protein dephosphorylates it.

In addition to these very exiting results, the new cell system can be used to further evolve the DHp-DX1 protein into a more efficient version that could also solve the *in vitro* detection problems. This can be possible by altering the cell system introducing genes important for cell survival that replace the reporter genes used previously, allowing for selection of more functional proteins that facilitate cell survival.

In vivo evolution strain selection

Construction of many different versions of the cell system were attempted. We constructed the versions shown in Figure 10 that include the cell lines JM_Kan_LVA, JM_Amp, JM_Amp_LVA, JM1012(pJM1,2(SacB)) and JM1012(pJM1,2(SacB_LVA)) with or without the addition of either the Taz, DHp-DX1, or the mutant DHp-DX1(H15Q). All cell lines were screened under different concentrations of the target toxin, whether this was kanamycin, ampicillin or sucrose (see materials and methods). Even though most of the strains failed to provide an effective selection because of direct interference of kanamycin and sucrose with the system[39,40,78], the ampicillin resistance strains showed good selection sensitivity. The cell line JM_Amp_LVA was especially effective in selecting the cells with an OmpR phosphorylase activity under semi-lethal conditions (50μM ampicillin; Figure 11). Interestingly, once more we could see that the fusion protein DHp-DX1 demonstrated half of the function observed for Taz, while the cells with the mutant DHp-DX1(H15Q) showed no effect.

In vivo evolution

Chemocompetent and electrocompetent cells of the JM_Amp_LVA strain including the plasmids pCL1920(DHp-DX1) and pCP20 (λRed) were transformed directly with the purified products of the random mutagenic PCR of the DHp-DX1 gene. The mutants were directly inserted in place of the original DHp-DX1 gene, through the λRed mediated homologous recombination. The selection of the cells was performed under semi-lethal conditions of ampicillin. This is an ongoing experiment.

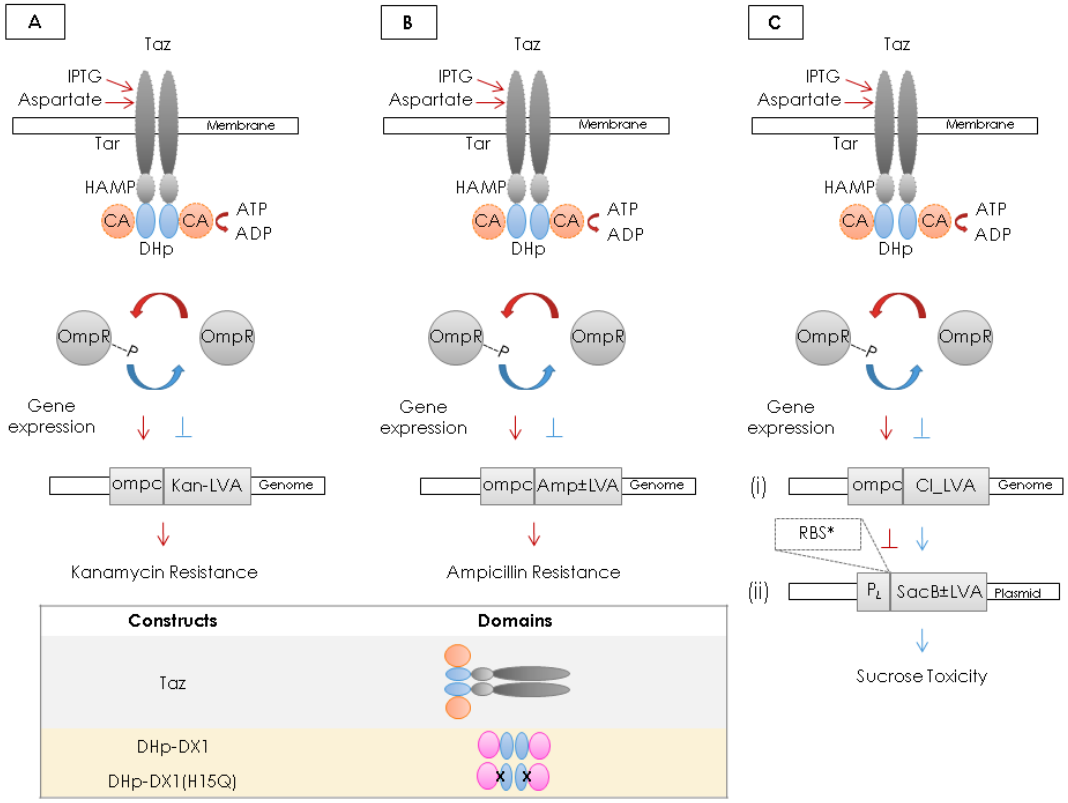


Figure 10. Construction of strains for *in vivo* evolution. (A) JM_Kan_LVA strain having the kanamycin resistance gene along with the destabilizing LVA tail in the chromosome and under control of the ompC promoter. **(B)** JM_Amp_LVA and JM_Amp strains having the ampicillin resistance gene with and without the LVA tail in the chromosome and under control of the ompC promoter. **(C)** JM1012(pJM1,2(SacB_LVA)) and JM1012(pJM1,2(SacB)) strains having the toxin in the presence of sucrose gene SacB with and without the LVA tail found in either the pJM1 or the pJM2 plasmid negatively controlled by the P_L promoter, which is repressed by the ompC controlled CI_LVA repressor. Both a weak or a strong RBS were tested (*).

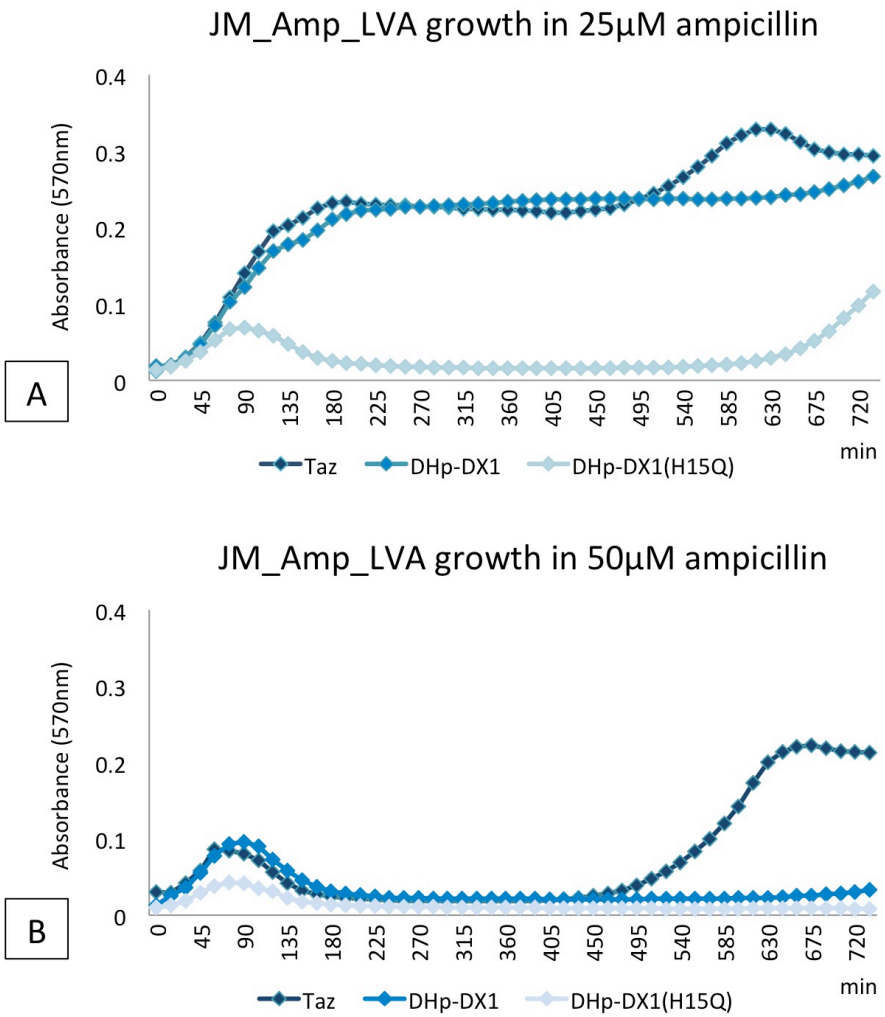


Figure 11. Cell growth of the JM_Amp_LVA stain under different ampicillin concentrations. The cells were cloned with either the protein Taz, DHp-DX1 or the histidine deficient mutant DHp-DX1(H15Q) and were grown in LB, at 37°C for 12h and in the presence of different ampicillin concentrations (0-100μM) to test the OmpR kinase selection ability of the system. **(A)** Growth curves with 25μM ampicillin. Taz and DHp-DX1 containing cells grow very similarly to each other, the mutant DHp-DX1(H15Q) fails to grow. **(B)** Growth curves with 50μM ampicillin. Only Taz containing cells show any growth. 50μM ampicillin can successfully select for improved OmpR kinase activity.

Conclusions

It is beyond any question that my protein DHp-DX1 activates the gene expression through the ompC promoter and that this effect is completely reversed when the target for phorsphorylation histidine on the DHp domain is mutated to glutamine. These results in combination with the observation of the OmpR phosphorylation *in vitro* at least once, strongly suggest that I have a functional protein that phosphorylates OmpR through a mechanism that involves the histidine, but has low efficiency. This low efficiency is also observed in the reduction of the binding affinity of the DHp domain when found in the DHp-DX1 fusion protein.

The strong ATPase activity of the DHp-DX1 protein described in the paper "Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase. " also poses more questions regarding the efficiency and function of the protein. Could this ATPase activity be a second function of the protein that is simply enhanced *in vitro,* suppressing a possible histidine phosphorylation to a point that is no longer detectable, or is the mechanism of function in the cell completely different and somehow depended on the ATPase functionality? The latter is quite a farfetched theory because the histidine mutation on the DHp domain does not affect the ATPase activity, but completely abolishes the *in vivo* observed gene expression controlled by the ompC promoter. In conclusion a dual functionality that shifts between a histidine kinase and an ATPase depending on the environment seems highly probable.

This multifunctionality along with the low efficiency observed for binding and phosphorylating the OmpR response regulator simulates perfectly the function

one would expect to find in a primordial enzyme. In parallel, the strong ATPase activity of the protein also matches with the first function one would expect from a precursor of a kinase. A function that is part of the final functionality of the kinase and still remains present in some form. Furthermore, the elusive nature of the DHp-DX1 protein allows us to keep following the path evolution follows to increase the specificity of every protein. *In vivo* evolution following the trajectory that benefits the kinase over an ATPase activity could lead to a more efficient enzyme that would allow the *in vitro* observation of the actual mechanism of function.

These findings highlight the evolutionary significance of the present study, which provides for the first time not only a proof of concept for the evolution of primordial enzymes in general, but also simulates the evolution of histidine kinases. Histidine kinases could indeed be the product of the initial combination of a protein that randomly bound ATP, like DX, with the DHp domain that can only be found in histidine kinases. Our findings also show that even if the CA domain can be found in many different proteins, it was still selected based on opportunity, since a protein analogous to it could replace it into a cell system observed for multiple strains. This last part is a clear representation of divergent evolution of two evolutionary independent proteins that can replace each other in a functional system.

This independent evolution led to the formation of two very different proteins in both sequence and structure that only share the ability to bind ATP (Figure 3). Strangely, however, the DX domain shares some common structural features with the CA domain. Using the structure of DX (Protein data bank ID: 3ltc) and the online structure comparison tool Dali[79], a number of CA homologs appeared to share structural similarities with the sandwich like fold of the α-helix (α1) and the 3 β-sheets (β1-3) of the DX protein. The detected structural similarities included the β4, β5, β7 sheets and the α2 helix of the Bergerat fold (Figure 3.). The CA domains found during this search were identified as the blue-light-activated histidine kinase 2 (4r3a-A), the sensor histidine kinase DESK (5ium-A, 5ium-B) and the domains found in serine/threonine-protein kinase BRSK1 and 2 (5iri-A, 5iri-B, 4yom-A) and the serine/threonine-protein kinase KCC4 (3osm-A) that are similar to the CA domain. This finding is strange because the DX α1 helix

is found on the opposite site of the ATP binding pocket, while it shows common structural characteristics with the ATP binding pocket of different CA domains. The α1 helix opposite to the ATP binding site better fits the distance between the β-sheets and α-helix of the CA domain but is not participating in the DX ATP binding pocket. The α2 helix neighboring the DX ATP binding site is more distant from the β-sheets. This longer distance along with the lack of a capping loop like the one found in the CA domain, fails to properly seal the ATP ligand leaving it unusually accessible and vulnerable to water. This is a critical factor leading to the *in vivo*-observed inefficient kinase activity, and also the strong *in vitro*-observed ATPase activity of the DHp-DX1 fusion protein.

Finally, the significance of this study includes enzyme design. We uncovered areas of the protein DHp-DX1 where we could focus a design effort to assist a histidine kinase function. A good place to start would be the ATP binding pocket on the DX domain aiming to further protect the ATP ligand from water molecules, closer simulating the ATP binding pocket of the CA domain. The same would apply if we would tighter fit the DHp and the DX domains by optimizing the interface between them by docking. Another area to improve could be the binding site of OmpR that seems affected in the fused protein. An effort like that, however, would be more of a guessing game since we lack structural information that could illuminate the structural differences, which might cause this lower OmpR affinity.

The same principle could also be applied in designing other enzyme functions even those that have not been explored by nature but can be based on simple chemical reactions. This combination of *in vitro* evolution, natural domain utilization or "domain recycling" and the final *in vivo* evolution can be used to develop new enzymes with novel functionality and also to facilitate novel enzyme design.

Closing remarks

"Nothing in biology makes sense except in the light of evolution". This famous quote of Theodosius Dobzhansky from 1973[80] could not possibly describe the present study more accurately. During this trip we took through evolutionary time we encountered multiple protein evolution stories that led to either divergent or convergent functional or structural outcomes. Each story is unique, yet so well explained through functional centered hypotheses of why things turn out the way they did. It is amazing how in biology everything seems to make sense; it is the only science that entertains thoughts of a higher purpose. It is almost disheartening to think that all this beauty is simply a combination of luck and diversity, which allows adaptability. However, we should try and learn this very important lesson from evolution, the same way we could learn it from philosophy as Alan Watts put it "In music, though, one doesn't make the end of the composition the point of the composition. If that were so, the best conductors would be those who played fastest; and there would be composers who only wrote finales. People [would] go to concerts only to hear one crashing chord—because that's the end. Same way in dancing—you don't aim at a particular spot in the room; that's where you should arrive. The whole point of the dancing is the dance…We thought [that] life by analogy was a journey, was a pilgrimage, which had a serious purpose at the end. And the thing was to get to that end. Success, or whatever it is, or maybe heaven after you're dead. But we missed the point the whole way along. It was a musical thing, and you were supposed to sing, or to dance, while the music was being played."[81].

Acknowledgments

First a of all I would like to thank **Prof. Dr. Andrei N. Lupas** for selecting me to be part of his department and trusted me with very exciting and challenging projects helping me to learn and to develop not only scientifically but also personally. Thanks to him I am more secure in myself that I could ever imagine possible, something that I will always keep close to my heart to remain the accountable, responsible and truthful person he helped me become.

Second in line I could only thank **Dr. Edgardo Alfredo Sepúlveda Sánchez-Hidalgo** for not only teaching me the secrets of microbiology and molecular biology, but for also standing next to me as a true friend or even a brother whenever I needed help.

I would like to address a very special thank you to **Prof. Dr. Volkmar Braun** and to **Prof. Dr. Joachim E. Schultz** for advising me and encouraging me through some very difficult times of my PhD, always making me feel recognized and giving me the positive energy to move forward.

I would also like to thank **Prof. .Dr. Karl Forchhammer** for happily accepting to be part of my examination and advisory committee, following my progress every year and providing me with valuable feedback, helping me to improve and to finish my studies.

Similarly, I would like to thank **Dr. Marcus Hartmann** for participating in my advisory committee and always show great interest in my work, providing a very creative and unique insight to my projects.

I would really like to thank the rest of our team leaders
Dr. Birte Hernandez-Alvarez
Dr. Jörg Martin
Dr. Murray Coles
and **Dr. Vikram Alva-Kullanja**
for always being available to provide their input, their advice and also their actual efforts and work to facilitate not only me but everyone in the department providing opportunities for transfer of knowledge and collaboration.

I would like to thank **Dr. Hongbo Zhu** for always being willing and available to help and teach other people, and for always generously and kindly recognizing other people's work and efforts.

Additionally, I thank **Dr. Tjeerd Dijkstra** for being a positive force in our department. He was always offering to help and stand by so many people, not only scientifically but also personally, making a huge difference in their work, showing that a positive working environment is not only possible but also necessary for the growth of young researchers and the future of academia.

I also thank from the bottom of my heart **Dr. Harshul Arora Verasztó** for scientifically

helping me during the kinetics and crystallography experiments, but also for being a true friend, a sister and the most reliable person I have ever encountered.

Next, I could only thank **Dr. Iuliia Boichenko (Iatsenko)**, for our scientific conversations that helped me think deeper and for her friendship and love I have enjoyed over the years. You are one of the most amazing people I have ever met.

I would like to address a special thank you to **Maria Logotheti** for all the scientific discussions, her help during the crystallization experiments and most of all for being a true friend that stood next to me during my good and bad moments, showing nothing but understanding and kindness.

I thank **Dr. Adrian Fuchs** for being a great friend of mine, for always being here when I needed him both scientifically and personally, and for always being in for a great party.

Furthermore, I would like to thank:
Anja Rau for all the help in the lab, all her protocols, the translation of my abstract and her amazing character and heart.
Athina Iliopoulou for reminding me of home, always caring for me and keeping me fed
Eugen Netz for being a positive force in the Universe
Dr. Joana-Maria Soares Pereira for being there for me, for never judging anyone and for the coffee breaks, of course!
Lorena Maldoner for her calm power, her care for me and her patience with my German
Mateusz Korycinski for his scientific help during sequence analysis and for his unconditional kindness
Mikel Martinez Goikoetxea Fernandez Zarranz for pipetting, checking the progress of my experiments, taking me for coffee brakes, feeding me when I would forget and co-founding my third company

I am also grateful to **Dr. Reinhard Albrecht** and **Kerstin Bär** for all the crystallization setups, as well as **Astrid Ursinus, Eva Hertle** and **Silvia Deiss** for all the laboratory support. We would be lost without you.

I would really like to thank **Friday** (a.k.a. Fabian Mörtter) for all the philosophical discussions, for the long dancing and movie nights, for making me run 10km and for never forgiving me for biding his time.

I would also like to thank **Dr. Andre Noll** and **Dr. Leonid Fedorov** for the psychological support and their advice (but also for the pirate talks and the drinking games).

I kindly thank my external collaborators **Dr. Thomas Haug, Marion Rauser** and **Jonathan Mohr** from the Isotope & Radiation Protection Laboratory of the University of Tübingen, for imaging the autoradiography results, as well as **Anita Schultz, Anubha Seth, Julia Grishin** and **Manuel Finkbeiner** from the department of Pharmaceutical Chemistry, University of Tübingen for welcoming me in their laboratory.

I would like to especially thank the late **Prof. Dr. Elisa Izaurralde** from the Department of Biochemistry from the Max-Planck Institute for Developmental Biology in Tübingen, for her support during the autoradiography experiments and during my time as a PhD representative. I will never forget the valuable time she so generously offered me along with her unconditional commitment to science and all students.

I would also like to extent my thanks to **Dr. Günther Muth** from the department of Microbiology/Biotechnology that is part of the Interfaculty Institute of Microbiology and Infection Medicine (IMIT) and the Eberhard Karls University in Tübingen for happily providing me with the lambda red recombinase plasmid pCP20.

In addition, I would like to express my appreciation to **Prof. Boris Macek** and **Dr. Mirita Franz-Wachtel** from the Proteome center of the University of Tübingen, for performing mass spectrometry trials and **Prof. Sung Kuk Lee** from the Ulsan National Institute of Science and Technology (UNIST, Korea) for providing us with the JM1012, and pJM1,2 constructs.

I warmly thank **Prof. Dr Constantinos Stathopoulos** for always being by my site from the beginning of my scientific career, for recognizing me as a person and as a scientist and for helping me to get to where I am today. You will always be a friend to me.

I thank **Dr. Anne Regnier-Vigouroux** for accepting me in her lab in the Department of Neurobiology, INSERM-ATV of the Cancer Research Center of Germany (DKFZ) the first time I was around. I will never forget that she took care of me, guided me and opened so many doors for me, always pushing me forward.

(By the way, if you are still reading you are awesome, or weird, or still hoping to be mentioned...)

Last but not least I would like to thank my partner, my family and my future **Georgios Liolios** who deserves this PhD as much as I do, since he suffered with me every step of the way without even getting to enjoy the fun parts of it. You are and will always be the best thing that ever happened to me.

I also thank:
My mother **Meropi Nikolaidou** and father **Efstathios Karamichalis** for not only making me pretty but for also raising me to be a kind person with humor. Thank you for always supporting my dreams, no matter how crazy the sounded.
My sister **Eleni Karamichali** for saving my life and challenging me to draw better.
My brother **Ilias Karamichalis** for pushing me to be the best I could by keeping me to the highest possible standards as the very poor role model that I am (you can do better!).
My niece **Meropi-Alkmini Chionidou** for being an inspiration (and hopefully become the first astronaut of the family!).

I here want to thank all my other friends in Germany and in Greece who step by step pushed me to become the person that I am today (in chronological order).

Theocharis Tsiligkiris
Kiriaki Savva
Dimitrios Bempekis
Maria Zikou
Katerina Lazidou
Constantina Filipou
Maria Poulada
Dr. Apostolos Menegakis
Nikolaos Pouladas

Thank you all for the dance.

References

1. Kitzmiller v. Dover Area School District - Wikipedia. Available at: https://en.wikipedia.org/wiki/Kitzmiller_v._Dover_Area_School_District#cite_note-4. (Accessed: 17th August 2018)
2. trythinkingnow. *Judgment Day: Intelligent Design On Trial (FULL VIDEO) NOVA*.
3. Darwin, C. *Life and Letters of Charles Darwin*. **2,** (London: John Murray, 1888).
4. Dawkins, R. *Climbing Mount Improbable*. **012,** (Penguin Books, 1996).
5. Williams, D. L. Light and the evolution of vision. *Eye* **30,** 173–178 (2016).
6. Elich, T. D. & Chory, J. Phytochrome: If It Looks and Smells Like a Histidine Kinase, Is It a Histidine Kinase? *Cell* **91,** 713–716 (1997).
7. Yeh, K.-C. & Lagarias, J. C. Eukaryotic Phytochromes: Light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 13976–13981 (1998).
8. Logsdon, J. M. & Doolittle, W. F. Origin of antifreeze protein genes: A cool tale in molecular evolution. *Proc. Natl. Acad. Sci.* **94,** 3485–3487 (1997).
9. Beeby, M. The bacterial flagellar motor and the evolution of molecular machines. *Biochemist* **40,** 4–9 (2018).
10. Alberts, B. *et al.* The RNA World and the Origins of Life. *Mol. Biol. Cell 4th Ed.* (2002).
11. Robertson, M. P. & Joyce, G. F. The Origins of the RNA World. *Cold Spring Harb. Perspect. Biol.* **4,** (2012).
12. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28,** 411–418 (2003).
13. Zhang, B. & Cech, T. R. Peptide bond formation by in vitro selected ribozymes. *Nature* **390,** 96–100 (1997).
14. Noller, H. F., Hoffarth, V. & Zimniak, L. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256,** 1416–1419 (1992).
15. Szathmáry, E. The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends Genet. TIG* **15,** 223–229 (1999).
16. Wilson, T. J. & Lilley, D. M. J. RNA catalysis—is that it? *RNA* **21,** 534–537 (2015).
17. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289,** 920–930 (2000).
18. Szilágyi, A., Kun, Á. & Szathmáry, E. Early evolution of efficient enzymes and genome organization. *Biol. Direct* **7,** 38 (2012).
19. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30,** 409–425 (1976).
20. Kim, H. S., Mittenthal, J. E. & Caetano-Anollés, G. MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics* **7,** 351 (2006).
21. Lazcano, A. & Miller, S. L. On the Origin of Metabolic Pathways. *J. Mol. Evol.* **49,** 424–431 (1999).
22. Caetano-Anollés, G., Kim, H. S. & Mittenthal, J. E. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci.* **104,** 9358–9363 (2007).

23. Newton, M. S., Arcus, V. L., Gerth, M. L. & Patrick, W. M. Enzyme evolution: innovation is easy, optimization is complicated. *Curr. Opin. Struct. Biol.* **48,** 110–116 (2018).
24. Tyzack, J. D., Furnham, N., Sillitoe, I., Orengo, C. M. & Thornton, J. M. Understanding enzyme function evolution from a computational perspective. *Curr. Opin. Struct. Biol.* **47,** 131–139 (2017).
25. Copley, S. D. Toward a Systems Biology Perspective on Enzyme Evolution. *J. Biol. Chem.* **287,** 3–10 (2012).
26. Wang, M. & Caetano-Anollés, G. The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World. *Structure* **17,** 66–78 (2009).
27. Wang, M., Yafremava, L. S., Caetano-Anollés, D., Mittenthal, J. E. & Caetano-Anollés, G. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17,** 1572–1585 (2007).
28. Lupas, A. Coiled coils: new structures and new functions. *Trends Biochem. Sci.* **21,** 375–382 (1996).
29. Hartmann, M. D. *et al.* α/β coiled coils. *eLife* **5,** (2016).
30. Lupas, A. N. & Bassler, J. Coiled Coils - A Model System for the 21st Century. *Trends Biochem. Sci.* **42,** 130–140 (2017).
31. Truebestein, L. & Leonard, T. A. Coiled-coils: The long and short of it. *Bioessays* **38,** 903–916 (2016).
32. Capra, E. J. & Laub, M. T. Evolution of two-component signal transduction systems. *Annu. Rev. Microbiol.* **66,** 325–347 (2012).
33. Dutta, R., Qin, L. & Inouye, M. Histidine kinases: diversity of domain organization. *Mol. Microbiol.* **34,** 633–640 (1999).
34. Cock, P. J. A. & Whitworth, D. E. Evolution of Prokaryotic Two-Component System Signaling Pathways: Gene Fusions and Fissions. *Mol. Biol. Evol.* **24,** 2355–2357 (2007).
35. Whitworth, D. E. & Cock, P. J. A. Evolution of prokaryotic two-component systems: insights from comparative genomics. *Amino Acids* **37,** 459–466 (2009).
36. Cai, S. J. & Inouye, M. EnvZ-OmpR Interaction and Osmoregulation in Escherichia coli. *J. Biol. Chem.* **277,** 24155–24161 (2002).
37. Ferris, H. U., Coles, M., Lupas, A. N. & Hartmann, M. D. Crystallographic snapshot of the Escherichia coli EnvZ histidine kinase in an active conformation. *J. Struct. Biol.* **186,** 376–379 (2014).
38. Forst, S., Delgado, J. & Inouye, M. Phosphorylation of OmpR by the osmosensor EnvZ modulates expression of the ompF and ompC genes in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 6052–6056 (1989).
39. Hu, W.-C., MacDonald, R., Oosthuizen, J. L. & Melanie van Soeren. Sub-Inhibitory Kanamycin Changes Outer Membrane Porin Ratios in Escherichia coli B23 by Increasing the Level of OmpC. *J. Exp. Microbiol. Immunol. JEMI* **15,** 7 (2011).
40. Lo, J., Tessa Van Tol, Stephanie Yeung & Kevin Zou. EnvZ is not essential for the upregulation of OmpC following treatment with sublethal kanamycin in Escherichia coli. *J. Exp. Microbiol. Immunol. J EMI* **18,** 5 (2014).
41. Smarajit Chakraborty, R. S. W., Leslie K. Morgan, J. Y. & Linda J. Kenney. Non-canonical activation of OmpR drives acid and osmotic stress responses in single bacterial cells.
42. Utsumi, R. *et al.* Activation of bacterial porin gene expression by a chimeric signal transducer in response to aspartate. *Science* **245,** 1246–1249 (1989).
43. Levskaya, A. *et al.* Synthetic biology: Engineering *Escherichia coli* to see light. *Nature* **438,** 441–442 (2005).
44. Michalodimitrakis, K. M., Sourjik, V. & Serrano, L. Plasticity in amino acid sensing of the chimeric receptor Taz. *Mol. Microbiol.* **58,** 257–266 (2005).

45. Lee, J. M., Lee, J., Kim, T. & Lee, S. K. Switchable gene expression in Escherichia coli using a miniaturized photobioreactor. *PloS One* **8,** e52382 (2013).
46. Ferris, H. U. *et al.* The Mechanisms of HAMP-Mediated Signaling in Transmembrane Receptors. *Structure* **19,** 378–385 (2011).
47. Mechaly, A. E., Sassoon, N., Betton, J.-M. & Alzari, P. M. Segmental helical motions and dynamical asymmetry modulate histidine kinase autophosphorylation. *PLoS Biol.* **12,** e1001776 (2014).
48. Bergerat, A. *et al.* An atypical topoisomerase II from Archaea with implications for meiotic recombination. *Nature* **386,** 414–417 (1997).
49. Campbell, E. A. *et al.* Crystal structure of the Bacillus stearothermophilus anti-sigma factor SpoIIAB with the sporulation sigma factor sigmaF. *Cell* **108,** 795–807 (2002).
50. Min, K. T., Hilditch, C. M., Diederich, B., Errington, J. & Yudkin, M. D. Sigma F, the first compartment-specific transcription factor of B. subtilis, is regulated by an anti-sigma factor that is also a protein kinase. *Cell* **74,** 735–742 (1993).
51. Harris, R. A. *et al.* A new family of protein kinases--the mitochondrial protein kinases. *Adv. Enzyme Regul.* **35,** 147–162 (1995).
52. Machius, M., Chuang, J. L., Wynn, R. M., Tomchick, D. R. & Chuang, D. T. Structure of rat BCKD kinase: nucleotide-induced domain communication in a mitochondrial protein kinase. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11218–11223 (2001).
53. Prodromou, C. *et al.* Identification and Structural Characterization of the ATP/ADP-Binding Site in the Hsp90 Molecular Chaperone. *Cell* **90,** 65–75 (1997).
54. Wigley, D. B., Davies, G. J., Dodson, E. J., Maxwell, A. & Dodson, G. Crystal structure of an N-terminal fragment of the DNA gyrase B protein. *Nature* **351,** 624–629 (1991).
55. Ban, C. & Yang, W. Crystal Structure and ATPase Activity of MutL: Implications for DNA Repair and Mutagenesis. *Cell* **95,** 541–552 (1998).
56. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410,** 715–718 (2001).
57. Simmons, C. R. *et al.* A Synthetic Protein Selected for Ligand Binding Affinity Mediates ATP Hydrolysis. *ACS Chem. Biol.* **4,** 649–658 (2009).
58. Smith, M. D. *et al.* Structural Insights into the Evolution of a Non-Biological Protein: Importance of Surface Residues in Protein Fold Optimization. *PLoS ONE* **2,** e467 (2007).
59. Paupe, V., Prudent, J., Dassa, E. P., Rendon, O. Z. & Shoubridge, E. A. CCDC90A (MCUR1) is a cytochrome c oxidase assembly factor and not a regulator of the mitochondrial calcium uniporter. *Cell Metab.* **21,** 109–116 (2015).
60. Vais, H. *et al.* MCUR1, CCDC90A, Is a Regulator of the Mitochondrial Calcium Uniporter. *Cell Metab.* **22,** 533–535 (2015).
61. Balczun, C. *et al.* Two adjacent nuclear genes are required for functional complementation of a chloroplast trans-splicing mutant from Chlamydomonas reinhardtii. *Plant J.* **43,** 636–648 (2005).
62. Zimmerly, S. & Semper, C. Evolution of group II introns. *Mob. DNA* **6,** (2015).
63. Surdo, P. L., Walsh, M. A. & Sollazzo, M. A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nat. Struct. Mol. Biol.* **11,** 382–383 (2004).
64. Audic, S., Lopez, F., Claverie, J. M., Poirot, O. & Abergel, C. SAmBA: an interactive software for optimizing the design of biological macromolecules crystallization experiments. *Proteins* **29,** 252–257 (1997).
65. Dutta, R., Yoshida, T. & Inouye, M. The critical role of the conserved Thr247 residue in the functioning of the osmosensor EnvZ, a histidine Kinase/Phosphatase, in Escherichia coli. *J. Biol. Chem.* **275,** 38645–38653 (2000).
66. McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random mutagenesis by error-prone PCR. *Methods Mol. Biol. Clifton NJ* **634,** 103–109 (2010).

67. Georgiou, G. & Valax, P. Expression of correctly folded proteins in Escherichia coli. *Curr. Opin. Biotechnol.* **7,** 190–197 (1996).
68. Gueguen, E. *et al.* Expression of a Yersinia pseudotuberculosis Type VI Secretion System Is Responsive to Envelope Stresses through the OmpR Transcriptional Activator. *PLoS ONE* **8,** (2013).
69. Papaneophytou, C. P. & Kontopidis, G. Statistical approaches to maximize recombinant protein expression in Escherichia coli: A general review. *Protein Expr. Purif.* **94,** 22–32 (2014).
70. Noguère, C., Larsson, A. M., Guyot, J.-C. & Bignon, C. Fractional factorial approach combining 4 Escherichia coli strains, 3 culture media, 3 expression temperatures and 5 N-terminal fusion tags for screening the soluble expression of recombinant proteins. *Protein Expr. Purif.* **84,** 204–213 (2012).
71. Cieśla, J., Frączyk, T. & Rode, W. Phosphorylation of basic amino acid residues in proteins: important but easily missed. *Acta Biochim. Pol.* **58,** 137–148 (2011).
72. Lapek, J. D., Tombline, G. & Friedman, A. E. Mass spectrometry detection of histidine phosphorylation on NM23-H1. *J. Proteome Res.* **10,** 751–755 (2011).
73. Potel, C. M., Lin, M.-H., Heck, A. J. R. & Lemeer, S. Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics. *Nat. Methods* **15,** 187 (2018).
74. Yoshida, T., Qin, L. & Inouye, M. Formation of the stoichiometric complex of EnvZ, a histidine kinase, with its response regulator, OmpR. *Mol. Microbiol.* **46,** 1273–1282 (2002).
75. Zhu, Y., Qin, L., Yoshida, T. & Inouye, M. Phosphatase activity of histidine kinase EnvZ without kinase catalytic domain. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 7808–7813 (2000).
76. Igo, M. M., Ninfa, A. J., Stock, J. B. & Silhavy, T. J. Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor. *Genes Dev.* **3,** 1725–1734 (1989).
77. Sourjik, V. & Berg, H. C. Receptor sensitivity in bacterial chemotaxis. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 123–127 (2002).
78. Wang, L. C., Morgan, L. K., Godakumbura, P., Kenney, L. J. & Anand, G. S. The inner membrane histidine kinase EnvZ senses osmolality via helix-coil transitions in the cytoplasm. *EMBO J.* **31,** 2648–2659 (2012).
79. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44,** W351–W355 (2016).
80. Dobzhansky, T. Nothing in Biology Makes Sense except in the Light of Evolution. in *The American Biology Teacher* **35,** 125–129 (National Association of Biology Teachers, 1973).
81. Myna. *Alan Watts - Life Is A Dance*.
82. Andersen, J. B. *et al.* New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl. Environ. Microbiol.* **64,** 2240–2246 (1998).
83. Miroux, B. & Walker, J. E. Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J. Mol. Biol.* **260,** 289–298 (1996).
84. Ibrahim, I. M., Puthiyaveetil, S. & Allen, J. F. A Two-Component Regulatory System in Transcriptional Control of Photosystem Stoichiometry: Redox-Dependent and Sodium Ion-Dependent Phosphoryl Transfer from Cyanobacterial Histidine Kinase Hik2 to Response Regulators Rre1 and RppA. *Plant Cell Biol.* 137 (2016). doi:10.3389/fpls.2016.00137

Appendix

Constructs

Table 2. Cell lines used in the present study.

Name	Characteristics	Source
ArcticExpress™, *E. coli*	LB/37-13oC	Agilent Technologies
AT142, *E. coli*	LB or M9 minimal/37oC	Michalogiannakis *et al.* (2005)[42,44,82]
BL21 gold, *E. coli* *	LB or M9 minimal /30-37oC	Novagen
C41, *E. coli*	LB/37oC	Miroux & Walker (1996)[83]
DH5AΛpir, *E. coli*	LB/37oC	In Lab
JM1012, *E. coli*	MG1655/PompC::cI-LVA ΔenvZ LB or M9 minimal/37oC	Ibrahim et al. (2016)[84]
JM_Amp, *E. coli*	JM1012/ PompC::Amp LB/37oC	Present Study
JM_Amp_LVA, *E. coli*	JM1012/ PompC::Amp-LVA LB/37oC	Present Study
JM_Kan_LVA, *E. coli*	JM1012/ PompC::Kan-LVA LB/37oC	Present Study
Rosetta 2 (DE3) pLysS**	LB/16-37oC	Novagen

*Used for the overexpression of the TRX fused DHp-CA, DHp-DX1, DHp-DX2, DX-DHp (including any of their mutants)
** Used for the overexpression of the TRX fused OmpR and DX

Table 3. Constructs used in the present study.

Name	Description	Method	Source
DHp-DX1,2 or DX-DHp°	DHp_{EnvZ}+DX Fusion Models	Synthesis	Eurofins Genomics
pETTRX(DHp-DX1,2 or DX-DHp)	Fusion Models plus N-term. Thierodoxin/His Tag*	PCR	Present Study
pETTRX(DHp-DX1(H15Q))	Fusion Models His15^Gln15 plus N-term. Thierodoxin/His Tag*	PCR	Present Study
pETTRX(DHp-CA)	Positive control (DHp_{EnvZ}-CA_{EnvZ})*	PCR	Present Study
pETTRX(OmpR)	Response regulator $OmpR_{EnvZ}$**	PCR	Present Study
pETTRX(DX)	Negative control DX*	PCR	Present Study
pETTRX(DHp-CA(T19Q))	Positive control (DHp_{EnvZ}-CA_{EnvZ}) Thr19^Gln19 *	PCR	Present Study
pCL1920	Negative control/Plasmid with no insert	-	Present Study
pCL1920(taz)	Chimeric Taz TCST system in the pCL1920 plasmid***	-	Michalogiannakis *et al.* (2005)[42,44,82]
pCL1920(Tar-DHp)	Control (Tar-HAMP-DHp_{EnvZ})***	-	Present Study
pCL1920(DHp-CA)	Control (DHp_{EnvZ}-CA_{EnvZ})***	-	Present Study
pCL1920(DHp)	Control (DHp_{EnvZ})***	-	Present Study
pCL1920(HDHp-CA)	Control ($HAMP_{AF1503}$-DHp_{EnvZ}-CA_{EnvZ})***	PCR	Present Study
pCL1920(HDHp)	Control ($HAMP_{AF1503}$-DHp_{EnvZ})***	PCR	Present Study
pCL1920(DHp-DX1,2 or DX-DHp)	Fusion Models***	-	Present Study
pCL1920(DHp-DX1(H15Q))	F2 fusion protein (DHp_{EnvZ}-natural linker-DX) histidine deficient mutant***	PCR	Present Study
pCL1920(Tar-HAMP-DHP-DX1,2)	Fusion Models with Tar sensor and $HAMP_{EnvZ}$ domain***	-	Present Study
pCL1920($HAMP_{AF1503}$-DHp-DX1,2 or $HAMP_{AF1503}$-DX-DHp)	$HAMP_{AF1503}$-Fusion Models***	-	Present Study
pET3a(gfp(LVA))	Reporter gene/destabilized GFP-LVA	-	Michalogiannakis *et al.* (2005)[42,44,82]
pET3a(GFP)	Reporter gene/wild type GFP	PCR	Present Study
pJM1,2(GFP)	Reporter gene/wild type GFP	-	Lee et al. (2013)[45]
pJM1(SacB)	Reporter gene/wild type SacB	PCR	Present Study
pJM1(SacB(LVA))	Reporter gene/SacB-LVA	PCR	Present Study
pETHis	N-term. His Tag*	-	In lab
pET28a-C(+)	C-term.His Tag*	-	Novagen
pCP20	λRed homologous recombinase vector	-	Provided by Dr. Günther Muth
pETTRX	N-term. Thierodoxin/His Tag*	-	In lab
pET-15b	No tag expression vector	-	Novagen

*4h induction at 37°C using 0.25 IPTG at OD equal to 0.6-0.8. DHp-DX1.2 proteins were expressed under 30°C
**16h induction at 16°C using 0.25 IPTG at OD equal to 0.6-0.8
***Overnight induction at 37°C using 0.25mM IPTG at OD equal to 0.3
°The DX-DHp ,DHp-DX1 and DHp-DX2 constructs can be found as F1, F2 and F3 in the stock collection.

Selected sequences

Vector:	**pCL1920**
Insert:	DHp-DX1
Organism:	*E. coli* DH5AΛpir
Selection:	Spectinomycin
Made by:	Ioanna

References and comments:
In silico designed protein containing DHp (Ferris *et al.* 2014) and DX (Simmons *et al.* 2010) domain in this order, with the linker found between DHp-CA domains in the EnvZ kinase (Ferris *et al.* 2014).

DNA-Sequence:
ATGACCATGATTACGCCAAGCTTGCAGTTAGCCGATGATCGCACCCTGTTAATGGCTGGTGTGAGTCATGATCTCCGTAC
ACCGCTCACCCGGATTCGCCTGGCAACCGAAATGATGAGCGAACAAGACGGCTATCTGGCAGAATCGATTAACAAAGATA
TTGAGGAGTGTAATGCCATCATCGAACAGTTTATCGACTATCTGCGTACTGGCCAGGAAATGCCGGGTGACAAGAAAACG
AATTGGCTGAAACGCATTTACCGTGTACGCCCATGTGTGAAATGCAAAGTTGCACCTCGTGACTGGAAAGTCAAGAACAA
ACACCTTCGGATTTACAACATGTGCAAAACGTGCTTTAACAACTCCATTGACATTGGTGACGATACCTATCATGGCCATG
TGGATTGGCTGATGTATGCGGATTCATAA

Protein-Sequence:
MTMITPSLQLADDRTLLMAGVSHDLRTPLTRIRLATEMMSEQDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPGDKKT
NWLKRIYRVRPCVKCKVAPRDWKVKNKHLRIYNMCKTCFNNSIDIGDDTYHGHVDWLMYADS
Changed to Q for DHp-DX1(H15Q) mutant

Is expressed: *In vivo* (*E. coli* AT142)

Product: PCR product derived from ordered gene from Eurofins.

Primers:

No	Name	Enzyme	%GC	TM	Sequence	Size
1	PCL_F2_F_HindIII	Phusion	56	62	TTATCCAAGCTTGCAGTTAGCCGATGATCGC	18
2	pCL1920F2_R_BamHI	Phusion	41	61	TTATCCGGATCCTTATGAATCCGCATACATCAGC	22

Restriction sites:
HindIII/BamHI

Vector:	**petTRX**
Insert:	DHp-DX1
Organism:	*E. coli* BL21 gold
Selection:	Kanamycin
Made by:	Ioanna

References and comments:
In silico designed protein containing DHp (Ferris *et al.* 2014) and DX (Simmons *et al.* 2010) domain in this order, with the linker found between DHp-CA domains in the EnvZ kinase (Ferris *et al.* 2014). The N-terminus TRX fusion stabilizes the protein during overexpression and purification and it can be removed by TEV cleavage.

DNA-Sequence:
ATGAGCGATAAAATTATTCACCTGACTGACGACAGTTTTGACACGGATGTACTCAAAGCGGACGGGGCGATCCTCGTCGA
TTTCTGGGCAGAGTGGTGCGGTCCGTGCAAAATGATCGCCCCGATTCTGGATGAAATCGCTGACGAATATCAGGGCAAAC
TGACCGTTGCAAAACTGAACATCGATCAAAACCCTGGCACTGCGCCGAAATATGGCATCCGTGGTATCCCGACTCTGCTG
CTGTTCAAAAACGGTGAAGTGGCGGCAACCAAAGTGGGTGCACTGTCTAAAGGTCAGTTGAAAGAGTTCCTCGACGCTAA
CCTGGCCGGATCTGGCAGTGGTTCTGGTCATCACCATCACCATCACTCCGCGGGTAGCGAGAATCTTTATTTTCAGGGCG
CCATGGGACAGTTAGCCGATGATCGCACCCTGTTAATGGCTGGTGTGAGTCATGATCTCCGTACACCGCTCACCCGGATT
CGCCTGGCAACCGAAATGATGAGCGAACAAGACGGCTATCTGGCAGAATCGATTAACAAAGATATTGAGGAGTGTAATGC
CATCATCGAACAGTTTATCGACTATCTGCGTACTGGCCAGGAAATGCCGGGTGACAAGAAAACGAATTGGCTGAAACGCA
TTTACCGTGTACGCCCATGTGTGAAATGCAAAGTTGCACCTCGTGACTGGAAAGTCAAGAACAAACACCTTCGGATTTAC
AACATGTGCAAAACGTGCTTTAACAACTCCATTGACATTGGTGACGATACCTATCATGGCCATGTGGATTGGCTGATGTA
TGCGGATTCATAA

Protein-Sequence:
MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLL
LFKNGEVAATKVGALSKGQLKEFLDANLAGSGSGSGHHHHHHSAGSENLYFQGAMGQLADDRTLLMAGVSHDLRTPLTRI
RLATEMMSEQDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPGDKKTNWLKRIYRVRPCVKCKVAPRDWKVKNKHLRIY
NMCKTCFNNSIDIGDDTYHGHVDWLMYADS
Changed to Q for DHp-DX1(H15Q) mutant

Is expressed: *In vivo* (*E. coli* BL21 gold)/0.6OD/0.25mM IPTG/2% Ethanol/4h/30oC. The expression levels were very good. Purification can be done in one step (Ni) in a basic environment (pH 8.5, 30mM Tris, 500mM NaCl), Dialyze in pH8,5, 30mM Tris, 200mM NaCl, 5mM 2-mercaptoethanol, 5mM citrate, 10μM $ZnCl_2$. Removing the TRX with TEV protease overnight in -4^{o}C in dialysis buffer.

Product PCR product from DHp-DX1/PCL1920

Primers

No	Name	Enzyme	%GC	TM	Sequence	Size
1	PET_F2_F_NcoI	Phusion	56	62	TTATACCATGGGACAGTTAGCCGATGATCGC	18
2	pCL1920F2_R_BamHI	Phusion	41	61	TTATCCGGATCCTTATGAATCCGCATACATCAGC	22

Restriction sites NcoI/BamHI

Vector:	**pCL1920**
Insert:	DHp-CA
Organism:	*E. coli* DH5AΛpir
Selection:	Spectinomycin
Made by:	Ioanna

References and comments:
Tar-HAMP domains were removed through PCR from the Taz construct (Michalidimitrakis *et al.* 2005)

DNA-Sequence

```
ATGACCATGATTACGCCAAGCTTGGCGGCTGGTGTTAAGCaACTGGCGGATGACCGCACGCTGCTGATGGCGGGGGTAAG
TCACGACTTGCGTACACCGGTGACGCGTATTCGCCTGGCGACCGAGATGATGAGCGAGCAGGATGGCTACCTGGCAGAAT
CGATCAATAAAGATATCGAGGAGTGCAATGCCATCATTGAGCAGTTTATCGACTACCTGCGCACCGGACAGGAGATGCCG
ATGGAAATGGCGGATCTCAACGCAGTACTCGGTGAGGTTATTGCTGCCGAAAGTGGCTATGAGCGGGAAATTGAAACCGC
GCTTTACCCCGGCAGCATTGAAGTGAAAATGCACCCGCTGTCGATCAAACGCGCGGTGGCGAATATGGTGGTCAACGCCG
CCCGTTACGGCAATGGCTGGATCAAAGTCAGCAGCGGGACGGAGCCGAATCGCGCCTGGTTCCAGGTGGAAGATGACGGT
CCGGGAATTGCGCCGGAACAACGTAAGCACCTGTTCCAGCCGTTTGTCCGCGGCGATAGTGCGCGCACCATTAGCGGCAC
GGGATTAGGGCTGGCAATTGTGCAGCGTATCGTGGATAACCATAACGGGATGCTGGAGCTTGGCACCAGCGAGCGGGGCG
GGCTTTCCATTCGCGCCTGGCTGCCAGTGCCGGTAACGCGGGCGCAGGGCACGACAAAAGAAGGGTAA
```

Protein-Sequence

```
MTMITPSLAAGVKQLADDRTLLMAGVSHDLRTPVTRIRLATEMMSEQDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPMEMADL
NAVLGEVIAAESGYEREIETALYPGSIEVKMHPLSIKRAVANMVVNAARYGNGWIKVSSGTEPNRAWFQVEDDGPGIAPEQRKHLF
QPFVRGDSARTISGTGLGLAIVQRIVDNHNGMLELGTSERGGLSIRAWLPVPVTRAQGTTKEG
Changed to Q for DHp-CA(T19Q) mutant
```

Is expressed? In vivo (*E. coli* AT142)

Product? PCR product

Primers

No	Name	Enzyme	%GC	TM	Sequence	Size
1	pCL1920CA_R_BamHI	Phusion	47	61	AGTGAATTCGAGCTCGGTA	19
2	pCL1920DHp_F_HindIII	Phusion	63	62	TTATCCAAGCTTGGCGGCTGGTGTTAAGC	16

Restriction sites:
HindIII/BamHI

Vector:	**petTRX**
Insert:	DC
Organism:	*E. coli* BL21 gold
Selection:	Kanamycin
Made by:	Ioanna

References and comments:
Tar-HAMP domains were removed through PCR from the Taz construct (Michalidimitrakis *et al.* 2005)
Restriction sites where altered through PCR to NcoI/BamHI in order to ligate the insert in petTRX plasmid.

DNA-Sequence

```
TTGTTTACTTTAAGAAGGAGATATACCATGAGCGATAAAATTATTCACCTGACTGACGACAGTTTTGACACGGATGTACT
CAAAGCGGACGGGGCGATCCTCGTCGATTTCTGGGCAGAGTGGTGCGGTCCGTGCAAAATGATCGCCCCGATTCTGGATG
AAATCGCTGACGAATATCAGGGCAAACTGACCGTTGCAAAACTGAACATCGATCAAAACCCTGGCACTGCGCCGAAATAT
GGCATCCGTGGTATCCCGACTCTGCTGCTGTTCAAAAACGGTGAAGTGGCGGCAACCAAAGTGGGTGCACTGTCTAAAGG
TCAGTTGAAAGAGTTCCTCGACGCTAACCTGGCCGGATCTGGCAGTGGTTCTGGTCATCACCATCACCATCACTCCGCGG
GTAGCGAGAATCTTTATTTTCAGGGCGCCATGGGGCAACTGGCGGATGACCGCACGCTGCTGATGGCGGGGGTAAGTCAC
GACTTGCGTACACCGCTGACGCGTATTCGCCTGGCGACCGAGATGATGAGCGAGCAGGATGGCTACCTGGCAGAATCGAT
CAATAAAGATATCGAGGAGTGCAATGCCATCATTGAGCAGTTTATCGACTACCTGCGCACCGGACAGGAGATGCCGATGG
AAATGGCGGATCTCAACGCAGTACTCGGTGAGGTTATTGCTGCCGAAAGTGGCTATGAGCGGGAAATTGAAACCGCGCTT
TACCCCGGCAGCATTGAAGTGAAAATGCACCCGCTGTCGATCAAACGCGCGGTGGCGAATATGGTGGTCAACGCCGCCCG
TTACGGCAATGGCTGGATCAAAGTCAGCAGCGGGACGGAGCCGAATCGCGCCTGGTTCCAGGTGGAAGATGACGGTCCGG
GAATTGCGCCGGAACAACGTAAGCACCTGTTCCAGCCGTTTGTCCGCGGCGATAGTGCGCGCACCATTAGCGGCACGGGA
TTAGGGCTGGCAATTGTGCAGCGTATCGTGGATAACCATAACGGGATGCTGGAGCTTGGCACCAGCGAGCGGGGCGGGCT
TTCCATTCGCGCCTGGCTGCCAGTGCCGGTAACGCGGGCGCAGGGCACGACAAAAGAAGGGTAA
```

Protein-Sequence

```
LFTLRRRYTMSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKY
GIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLAGSGSGSGHHHHHHSAGSENLYFQGAMGQLADDRTLLMAGVSH
DLRTPLTRIRLATEMMSEQDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPMEMADLNAVLGEVIAAESGYEREIETAL
YPGSIEVKMHPLSIKRAVANMVVNAARYGNGWIKVSSGTEPNRAWFQVEDDGPGIAPEQRKHLFQPFVRGDSARTISGTG
LGLAIVQRIVDNHNGMLELGTSERGGLSIRAWLPVPVTRAQGTTKEG
```

Changed to Q for DHp-CA(T19Q) mutant

Is expressed: In vivo (*E. coli* BL21 gold)/0.6OD/0.25mM IPTG/37oC/4h

Product: PCR product

Primers:

No	Name	Enzyme	%GC	TM	Sequence	Size
1	pCL1920CA_R_BamHI	Phusion	47	61	AGTGAATTCGAGCTCGGTA	19
2	pET3a/DC_F_NcoI	Phusion	63	61	TTATCCATGGGGCAACTGGCGGATGACC	16

Restriction sites BamHI/NcoI

Construct Data Sheet	
Name:	JM_Amp_LVA
Made by:	Ioanna Karamichali
Construction:	JM1012 /PompC::cI-LVA ΔenvZ
Vector:	pCL1920/DHp-DX1 (Spectinomycin)
Selection:	Ampicillin (semi-lethal)
Comments:	Strain for *in vivo* evolution with Ampicillin selection
Publication:	-

References and comments:
The ampicillin resistance gene was cloned after the OmpC promoter in the genome.

Cells survive in 30μM ampicillin but not in 50μM where cells otherwise survive if pCL1920/Taz is used instead.

Product: PCR product

Primers:

No	Name	Enzyme	%GC	Tm(oC)	Sequence	Size
1	Amp_F1_inpET3a	Phusion	43	58	ATATGTATCCGCTCATGAGAC	21
2	Amp_R1	Phusion	45	60	TTAAGCTACTAAAGCGTAGTTTTCGTCGTTTGCTGCCCAATGCTTAATCAGTGAGG	20
3	Amp_F2	Phusion	36	60	TGGCATAAAAAAGCAAATAAAGGCATATAACAGAGGGTTAATAACATGAGTATTCAACATTTCCGTG	22
4	Amp_LVA_R2	Phusion	33	57	GCAGGCCCTTTGTTCGATATCAATCGAGATTAGAACTGGTAAACCAGACCTTAAGCTACTAAAGCGTAGTTTTC	24

Curriculum Vitae

Ioanna Karamichali
ikaramichali@gmail.com

WORK EXPERIENCE

Max-Planck Institute for Developmental of Biology, Tübingen, Germany **2014-Present**
PhD candidate, Department of Protein Evolution
Supervisor: Prof. Dr. Andrei Lupas

Freelance project developer, Thessaloniki, Hellas **2013-2014**
Communication Officer, Assistant manager of web applications
Projects: Creative cables, Dontsos

Office of public relations, Biomedical Research Foundation of the Academy of Athens, Hellas **2011-2012**
Assistant manager of organization and support of scientific conferences
Supervisor: Dimitra Tsouropli

X-Cell Center, Athens, Hellas **2010-2011**
Assistant editor of clinical studies
Supervisor: Dimitrios Dimitriadis

C. Constantinou, private Microbiological practice, Thessaloniki, Hellas **2008-2009**
Quality assurance manager of biochemical tests and bioanalyst
Supervisor: Constantinos Constantinou, MD

EDUCATION

Department of Biology, University of Athens, Hellas **2010-2012**
MSc in Microbial Biotechnology

Department of Life Sciences, University of Thessaly, Hellas **2003-2008**
Diploma in Biochemistry and Biotechnology

RESEARCH EXPERIENCE

Max-Planck Institute for Developmental of Biology, Dep. of Protein Evolution, Tübingen, Germany **2014-Present**
PhD candidate
Project titles:
- The Classification of a New Head-Neck-Stalk-Anchor Protein Superfamily.
- Designing a Novel Histidine Kinase.

Supervisor: Prof. Dr. Andrei Lupas

Department of Plant Breeding and Biometry, Agricultural University of Athens, Hellas **2011-2012**
Postgraduate research associate
Project title:
- Non-Transgenic Molecular Vaccines for the Reinforcement of the Antiviral Resistance in Plants.

Supervisors: Prof. Dr. George Skaracis, Dr. Andreas E. Voloudakis

Division of Biotechnology, Biomedical Research Foundation of the Academy of Athens, Hellas **2010-2012**
Postgraduate master student
Master thesis title:
- Fusion – Fission Events During the Evolution of Human Bacterial Pathogens.

Supervisors: Prof. Dr. Amalia D. Karagkouni, Dr. Sofia Kossida

Department of Neurobiology, INSERM-ATV, Cancer Research Center of Germany (DKFZ), Heidelberg, Germany **2007 (March-June)**
Undergraduate student
Senior thesis title:
- The Ceramide's Role in the Glioma Cell Death Pathway Induced by Microglia Secreted Factors.

Supervisors: Prof. Dr. Jean Rommelaere, Dr. Anne Regnier-Vigouroux, Prof. Dr. Constantinos Stathopoulos

Department of Biochemistry and Biotechnology, University of Thessaly, Larissa, Hellas **2006-2007**
Undergraduate part time trainee
Supervisors: Prof. Dr. Constantinos Stathopoulos, Dr. George Zacharioudakis

GRANTS

- Research grant by the program AVIVA from the National Center for Research (CNRS). **2011-2012**
- Travel scholarship granted by the Cancer Research Center of Germany (DKFZ). **2007 (March-June)**

CONFERENCES

- **Karamichali I**, Adlakha J, Hartmann MD, Hernandez Alvarez B, Lupas AN (2017) MempromCC: A new family of membrane-attached coiled coils in prokaryotes and mitochondria. *Alpbach Workshop: Coiled-Coil, Fibrous Repeat Proteins*, 3-8 September, Austria.
- **Karamichali I,** Diallinas G, Karagouni AD, Kossida S (2011) Fusion – fission events during the evolution of human bacterial pathogens. *Conference of the Biomedical Research Foundation of the Academy of Athens*, 24-25 June, Hellas.

PUBLICATIONS

- **Karamichali I**, Sepulveda E, Zhu H, Schultz JE, Coles M, Lupas AN. Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a de novo evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase. *(In preparation)*
- Adlakha J, **Karamichali I**, Sangwallek J1, Deiss S, Hipp K, Coles M, Hartmann MD, Lupas AN, Hernandez Alvarez B. MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles. *(Under submission)*
- Hartmann MD, Mendler CT, Bassler J, **Karamichali I**, Ridderbusch O, Lupas AN, Hernandez Alvarez B. (2016) α/β coiled coils. *Elife* (doi: 10.7554/eLife.11861)
- **Karamichali I**, Koumandou VL, Karagouni AD, Kossida S (2014), Frequent gene fissions associated with human pathogenic bacteria. *Genomics* (doi: 10.1016/j.ygeno.2014.02.001)

COURSES

- **Bioinformatics for Biochemists**, Max-Planck Institute, 1 week, Introduction to computational biology, sequence analysis and structure prediction. **2014**
- **Pymol**, Max-Planck Institute, 1 day, Protein structure visualization software **2015**
- **Python Programming,** Max-Planck Institute, 2 courses, 2 weeks each, Bioinformatics using Python **2015,2016**
- **Rosetta**, Max-Planck Institute, 3 days, Protein structure prediction and design software **2016**
- **Unix**, Max-Planck Institute, 1 week, Unix usage for beginners. **2016**
- **Scientific writing**, Max-Planck Institute, 2 courses, 4 weeks in total **2015,2018**
- **Career planning**, Max-Planck Institute, 3 courses, 2 weeks in total **2017,2018**
- **MedTech startup innovation course**, MedTech Startup school, University of Tübingen, 100 days **2017**

TECHNIQUES AND OTHER SKILLS

- Protein sequence analysis (Blast, Clustal, MPI Bionformatics toolkit, Python)
- Protein structure modeling (Modeller, Rosetta, PyMol, Javascript, Python)
- Cell culture (mammalian, bacterial cells)
- Gene/ Plasmid/ Strain design (PCR)
- Gene cloning (bacterial, chemical or electrocompetent cells)
- Protein expression/purification (bacterial cells, nickel chromatography)
- *In vivo*/*in vitro* protein functional characterization (colorimetric, fluorescence or radioactivity based, in gel, toxicity based)
- *In vivo* evolution
- Protein biophysical characterization (size exclusion chromatography, circular dichroism, native/SDS/isoelectric focusing)
- Other techniques (flow cytometry, elisa, western blot, autoradiography)

TECHNICAL SKILLS

Languages:

- English (Business fluent)
- German (B1)
- Hellenic (Mother language)

Computer skills:

- Unix working skills (Intermediate)
- Perl and Python programming (Intermediate)
- Windows working skills (Excellent)

OTHER ACTIVITIES

- Project leader during the MedTech Startup School program of the University of Tübingen. **2017 (June-October)**
- Writer for the science-focused web Magazine NeuroMag. **2017-Present**
- Founding member of the Sustainability PhD Initiative of the University of Tübingen. **2017-Present**
- Manager of the Student Life web page of the Max-Planck Institute of Developmental Biology. **2015-2017**
- PhD student representative. **2015-2017**
- Organization of the PhD Retreat of the Max-Planck Institute of Developmental Biology. **2015-2016**

eLIFE elifesciences.org

RESEARCH ARTICLE

α/β coiled coils

Marcus D Hartmann, Claudia T Mendler†, Jens Bassler, Ioanna Karamichali, Oswin Ridderbusch‡, Andrei N Lupas*, Birte Hernandez Alvarez*

Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany

Abstract Coiled coils are the best-understood protein fold, as their backbone structure can uniquely be described by parametric equations. This level of understanding has allowed their manipulation in unprecedented detail. They do not seem a likely source of surprises, yet we describe here the unexpected formation of a new type of fiber by the simple insertion of two or six residues into the underlying heptad repeat of a parallel, trimeric coiled coil. These insertions strain the supercoil to the breaking point, causing the local formation of short β-strands, which move the path of the chain by 120° around the trimer axis. The result is an α/β coiled coil, which retains only one backbone hydrogen bond per repeat unit from the parent coiled coil. Our results show that a substantially novel backbone structure is possible within the allowed regions of the Ramachandran space with only minor mutations to a known fold.

DOI: 10.7554/eLife.11861.001

***For correspondence:** andrei.lupas@tuebingen.mpg.de (ANL); birte.hernandez@tuebingen.mpg.de (BHA)

Present address: †Nuklearmedizinische Klinik und Poliklinik, Klinikum rechts der Isar, Technische Universität München, Munich, Germany; ‡Vossius and Partner, Siebertstraße, Germany

Competing interests: The authors declare that no competing interests exist.

Funding: See page 21

Received: 24 September 2015
Accepted: 14 January 2016
Published: 15 January 2016

Reviewing editor: Mingjie Zhang, Hong Kong University of Science and Technology, China

Introduction

α-Helical coiled coils are ubiquitous protein domains, found in a wide range of structural and functional contexts (***Lupas, 1996***). They were the first protein fold described in atomic detail (***Crick, 1953b***) and are also the only one whose backbone structure can be computed with parametric equations (***Crick, 1953a***), placing them at the forefront of protein design efforts (***Huang et al., 2014***; ***Joh et al., 2014***; ***Thomson et al., 2014***; ***Woolfson, 2005***).

The structure of coiled coils is understood at a level unrivaled by any other fold. They consist of at least two α-helices, wound into superhelical bundles and held together by a mostly hydrophobic core. In their most prevalent form they follow a heptad sequence repeat pattern. The seven positions in a heptad are labeled *a* – *g*, where positions *a* and *d* are oriented towards the core of the bundle and are thus mostly hydrophobic. Beyond the heptad repeat, a range of other periodicities is accessible to coiled coils, which is only restrained by the periodicity of the unperturbed α-helix (***Gruber and Lupas, 2003***). This restraint is responsible for the supercoiling of the bundle: As an ideal, straight α-helix has a periodicity of about 3.63 residues per turn, the heptad coiled coil has a left-handed twist to reduce the periodicity to 3.5 residues per turn with respect to the bundle axis. In hendecad coiled coils, the situation is reversed: 11 residues are accommodated in 3 helical turns, resulting in 11/3 = 3.67 residues per turn. As this is slightly above 3.63, hendecads are slightly right-handed. With the periodicity of pentadecad coiled coils, 15/4 = 3.75 residues per turn, right-handedness is as pronounced as left-handedness is in heptad coiled coils.

Many naturally occurring coiled coils contain transitions between segments of different periodicity (***Alvarez et al., 2010***; ***Hartmann et al., 2014***) or harbor discontinuities that retain the α-helical structure, but perturb the periodicity locally (***Parry, 2014***). The best understood discontinuities are insertions of 3 or 4 residues, which are close to the periodicity of 3.63 of α-helices (***Brown et al., 1996***; ***Hicks et al., 2002***; ***Lupas and Gruber, 2005***). The insertion of 3 residues is termed a stammer, the insertion of 4 residues a stutter. With 3 residues being less than one full turn of a helix, stammers lead to a local decrease in periodicity and an increase of left-handedness. Stutters have the opposite effect. Inserted into a heptad coiled coil, a stutter can locally extend one heptad to form a hendecad

eLife digest Proteins are made up of building blocks called amino acids. Groups of amino acids within the protein can then fold into three-dimensional shapes, one of the most common being a helical structure known as an α-helix. Two or more α-helices may be wound around each other to form a bundle called a coiled coil, which is found in many proteins. Each complete turn of an α-helix contains a set number of amino acids, but the number of amino acids in the turns of a coiled coil can vary. The most common pattern in a coiled coil has 7 amino acids over two turns, which is known as a heptad repeat.

When amino acids are added into or deleted from the heptad repeats, the number of amino acids in the turns of a coiled coil changes. However, it cannot increase too far beyond the number of amino acids in each turn of a normal α-helix because there is a limit to the amount of coiling that the helices can tolerate. Many naturally occurring coiled coils have regions where the overall α-helical structure is retained, even though there are small sections where the number of amino acids in a turn is disrupted. This may be due to insertions of small numbers of amino acids. Although the impact of some insertions (e.g. three or four at a time) has been studied, the effect of inserting other amounts of amino acids was not clear.

Hartmann et al. investigated what would happen when two or six amino acids were inserted into the heptad repeats of a coiled coil within a protein from bacteria. These numbers of amino acids have been predicted to cause the greatest strain on the coiled coil structure. The experiments show that inserting these numbers of amino acids caused so much strain that the three α-helices making up the coiled coil break apart and refold into a completely different type of structure called a β-strand. The three short β-strands then associate into a triangular structure that Hartmann et al. named a β-layer.

Further experiments showed that inserting the same numbers of amino acids into the heptad repeats of other coiled coil proteins also resulted in the formation of β-layers. Hartmann et al.'s findings suggest that the alternating α-helix and β-strand structures may help to make the proteins stronger and enable to carry out more versatile roles in cells.

DOI: 10.7554/eLife.11861.002

(7 + 4 = 11 -> 11/3) or, being delocalized over multiple heptads, lead to even higher periodicities like 18 residues over 5 turns (7 + 7 + 4 = 18 -> 18/5). Other periodicities can be brought about by the insertion of multiple stammers or stutters (e.g. 7 + 4 + 4 = 15 -> 15/4). These relationships are illustrated in ***Figure 1***, which shows the effects on coiled-coil periodicity resulting from consecutive insertions of stammers (blue lines) and stutters (green lines), and from their progressive delocalization (red lines).

However, there are limits to the periodicities coiled coils can assume, imposed by the degree of supercoiling the constituent helices can tolerate. The insertion of a stammer into a heptad coiled coil, leading locally to a periodicity of 10/3 = 3.33, was predicted to cause an overwinding of the helices (***Brown et al., 1996***). We could verify this experimentally: the structure of a stammer showed that the local overwinding introduced sufficient strain to cause the formation of a short 3_{10}-helical segment (***Hartmann et al., 2009***). We therefore assume that 3.33 (10/3) residues per turn mark the lower limit for periodicities. As this is about 0.3 residues per turn less than the periodicity of a perfectly straight helix, one might expect the upper limit at a periodicity of about 3.9. In fact the vast majority of known coiled-coil structures deviating from the heptad repeat have periodicities higher than 3.5 and the most extreme example is found in the trimeric autotransporter YadA, which has a local periodicity of 3.8 (19/5) (***Alvarez et al., 2010***).

In contrast to stammers and stutters, accommodating insertions of 1 or 5 residues is more demanding for the bundle. According to ***Figure 1*** they have to be delocalized over more than one heptad, as periodicities of 4.0 ((7+1)/2) or 2.66 ((7+1)/3) do not fall into the accessible range, and neither do 2.5 (0+5/2), 4.0 ((7+5)/3) or 3.0 ((7+5)/4). To retain α-helical structure, both insertions of 1 and 5 residues have to be delocalized over at least two heptads, leading to periodicities of 3.75 (15/4) and 3.8 (19/5), respectively. Interestingly, these periodicities can also be brought about by the insertion of 2 (15/4) and 3 (19/5) consecutive stutters. Alternatively, insertions of 1 residue (skip

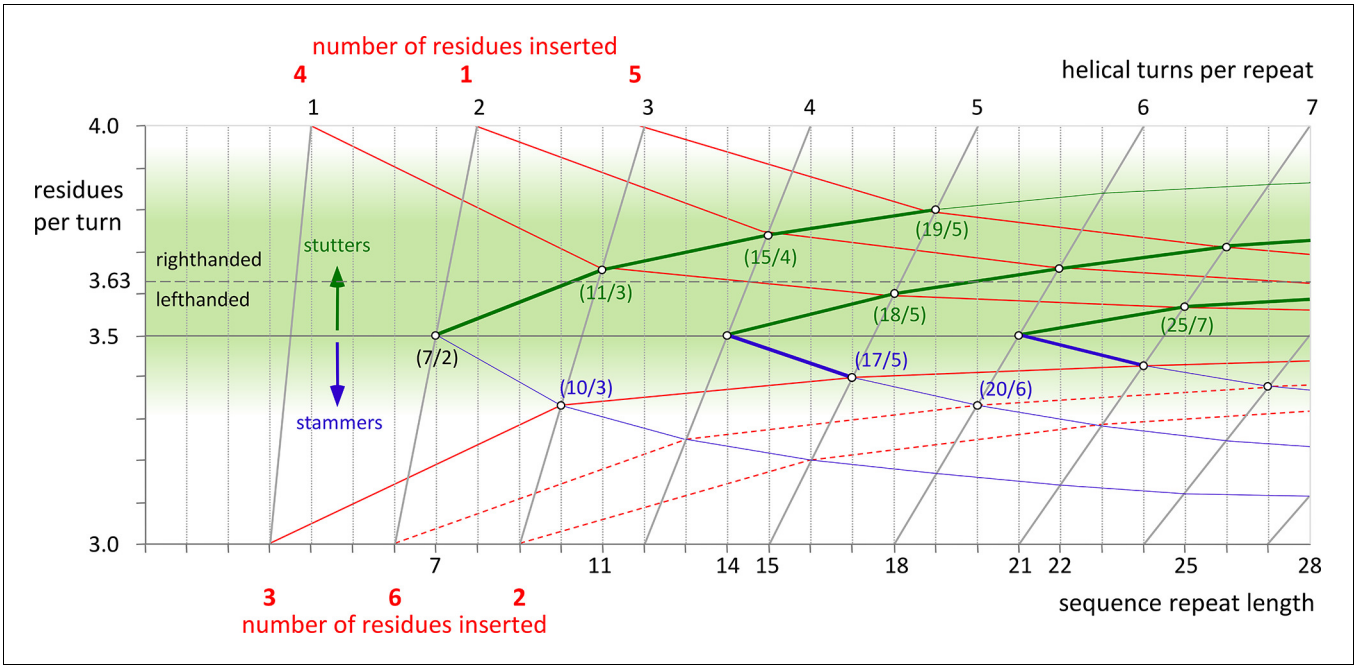


Figure 1. Transitions in periodicity caused by insertions of one to six residues into the heptad repeat. The green area marks the estimated boundaries of periodicities accessible to α-helical coiled coils. It is centered around the periodicity of unperturbed α-helices, about 3.63 residues per turn. Higher values than 3.63 lead to right-handed and lower values to left-handed supercoiling. The effects of consecutive insertions of stammers (3 residues) or stutters (4 residues) into a heptad pattern are shown by blue and green lines, respectively. The red lines correspond to the insertion of 1 to 6 residues into the heptad periodicity and their progressive delocalization over neighboring heptads. For example, an insertion of 4 residues is accommodated as 11 residues over 3 turns (11/3), when delocalized over one heptad, or as 18/5, when delocalized over two. Insertions of 1 or 5 residues have to be delocalized over two heptads, resulting in periodicities of 15/4 or 19/5 (which could also be brought about by consecutive stutters – following the green line from 7/2 over 11/3 over 15/4 to 19/5). Insertions of 3 can be accommodated as 10/3, at the very edge of the green area, although in the known examples the α-helices are distorted due to the strong left-handed supercoiling which could be avoided by further delocalization. For insertions of 2 or 6 residues (dashed lines) a strong delocalization would be required to reach the green lawn of accessible periodicities. However, for all constructs in this paper, this is not observed. Via the formation of *β*-layers these insertions sustain the heptad periodicity as unperturbed as possible.
DOI: 10.7554/eLife.11861.003

residues) can be accommodated by the local formation of a π-turn in the α-helix, leaving the remaining coiled coil largely unperturbed (***Lupas, 1996***).

Still missing for a complete picture of coiled-coil periodicities is the understanding of insertions of 2 and 6 residues, which should cause the greatest strain on α-helical geometry. We find that they indeed break the α-helices to form short β-strands, which associate into a triangular supersecondary structure we name the *β*-layer. *β*-Layers are found, also repetitively, in natural coiled coils, where they form regular fibers with alternating α- and β-structure, a protein fold that has not been described so far.

Results and discussion

A β-layer in the coiled-coil stalk of Actinobacillus OMP100

We have a long-standing interest in trimeric autotransporter adhesins (TAA), fibrous proteins of the Gram-negative bacterial surface (***Bassler et al., 2015***; ***Hartmann et al., 2012***; ***Hoiczyk et al., 2000***; ***Szczesny and Lupas, 2008***), whose domains we routinely fuse to stabilizing adaptor coiled coils for biochemical and biophysical study (***Deiss et al., 2014***; ***Hernandez Alvarez et al., 2008***). In the process, we have repeatedly gained insights into aspects of coiled-coil structure (***Alvarez et al., 2010***; ***Grin et al., 2014***; ***Hartmann et al., 2012***; ***2009***; ***Leo et al., 2011***), such as for example into a recurrent polar motif of the hydrophobic core (the N@*d* layer), in which asparagines in position *d* of the core coordinate anions at their center (***Hartmann et al., 2009***). As part of that study, we identified a putative TAA in *Actinobacillus actinomycetemcomitans*, OMP100, which carries insertions of 2 and

of 3 residues within the heptad repeats of its stalk. The insertion of 2 residues extends a heptad to the 9-residue motif IENKADKAD and occurs between three N-terminal and two C-terminal heptads carrying N@*d* layers; the insertion of 3 residues is directly downstream. This observation was highly puzzling, since the heptad register of the protein could be assigned with great confidence, leaving no doubt that an insertion of 2 residues had occurred, but this insertion could not be explained by coiled-coil theory. For structural characterization, we therefore expressed residues 133Q-198K, covering the two insertions and the five N@*d* layers, fused N- and C-terminally to the trimeric form of the GCN4 leucine zipper, GCN4-pII (***Table 1***). The construct yielded a typical α-helical CD spectrum and, upon heating, unfolded cooperatively with a transition midpoint at 91˚C. We obtained crystals in space group C2 that diffracted to a resolution of 2.3 Å, with one symmetric trimer in the asymmetric unit. The structure showed a continuous heptad coiled coil with two discontinuities (***Figure 2***). As expected, the insertion of 3 residues C-terminal to the N@*d* layers led to the formation of a decad, with a short 3_{10}-helical segment, as for the stammer we had described previously (***Hartmann et al., 2009***).

However, the insertion of 2 residues led to a sharp break in the coiled coil: In the middle of the IENKADKAD motif, the three chains of the trimer cross each other to form a triangular plane

Table 1. Sequences of constructs and protein buffer composition.

Construct	Protein sequence	Final buffer
OMP100	(GCN4-pII)$_N$-**IQNVDVR** **STENAAR** **SRANEQK** **IAENKKA** **IENKADKAD** **VEKNRAD** **IAANSRA** **IATFRSSSQN** **IAALTTK**-(GCN4pII)c-KLHHHHHH	20 mM Tris pH 7.5, 400 mM NaCl, 5% Glycerol
Tcar0761	(GCN4-N16V)$_N$-**ITLMQAN** **---MATKDD** **LARMATKDD** **IANMATKDD** **IANMATKDD** **IAKLDVK** **IENLNTK**-(GCN4-N16V)c-GSGHHHHHH	20 mM MOPS pH 7.2, 500 mM NaCl, 5% Glycerol, 2 M Urea
T6	(6xH-TEV)-(GCN4-N16V)$_N$-**MATKDD**-(GCN4-N16V)c	20 mM HEPES pH 7.4, 50 mM NaCl, 5% Glycerol, 1 M Urea
T9	(6xH-TEV)-(GCN4-N16V)$_N$-**MATKDDIAN**-(GCN4-N16V)c	20 mM HEPES pH 7.4, 50 mM NaCl, 5% Glycerol, 1 M Urea
A6	(6xH-TEV)-(GCN4-N16V)$_N$-**IENKAD**-(GCN4-N16V)c	20 mM HEPES pH 7.4, 50 mM NaCl, 5% Glycerol, 1 M Urea
A7	(6xH-TEV)-(GCN4-N16V)$_N$-**IENKKAD**-(GCN4-N16V)c	20 mM HEPES pH 7.4, 50 mM NaCl, 5% Glycerol, 1 M Urea
A9	(6xH-TEV)-(GCN4-N16V)$_N$-**IENKADKAD**-(GCN4-N16V)c	20 mM HEPES pH 7.4, 50 mM NaCl, 5% Glycerol, 1 M Urea
A9b	(6xH-TEV)-(GCN4-N16V)$_N$-**IANKEDKAD**-(GCN4-N16V)c	20 mM HEPES pH 7.5, 50 mM NaCl, 10% Glycerol, 1 M Urea

(GCN4-pII)N	MKQIEDKIEEILSKIYHIENEIARIKKL
(GCN4-pII)C	MKQIEDKIEEILSKIYHIENEIARIKKLI
(GCN4-N16V)N	MKQLEMKVEELLSKVYHLENEVARLKKL
(GCN4 N16V)C	MKQLEWKVEELLSKVYHLENEVARLKKLV
(6xH-TEV)	MKHHHHHHPMSDYDIPTTENLYFQGH

DOI: 10.7554/eLife.11861.004

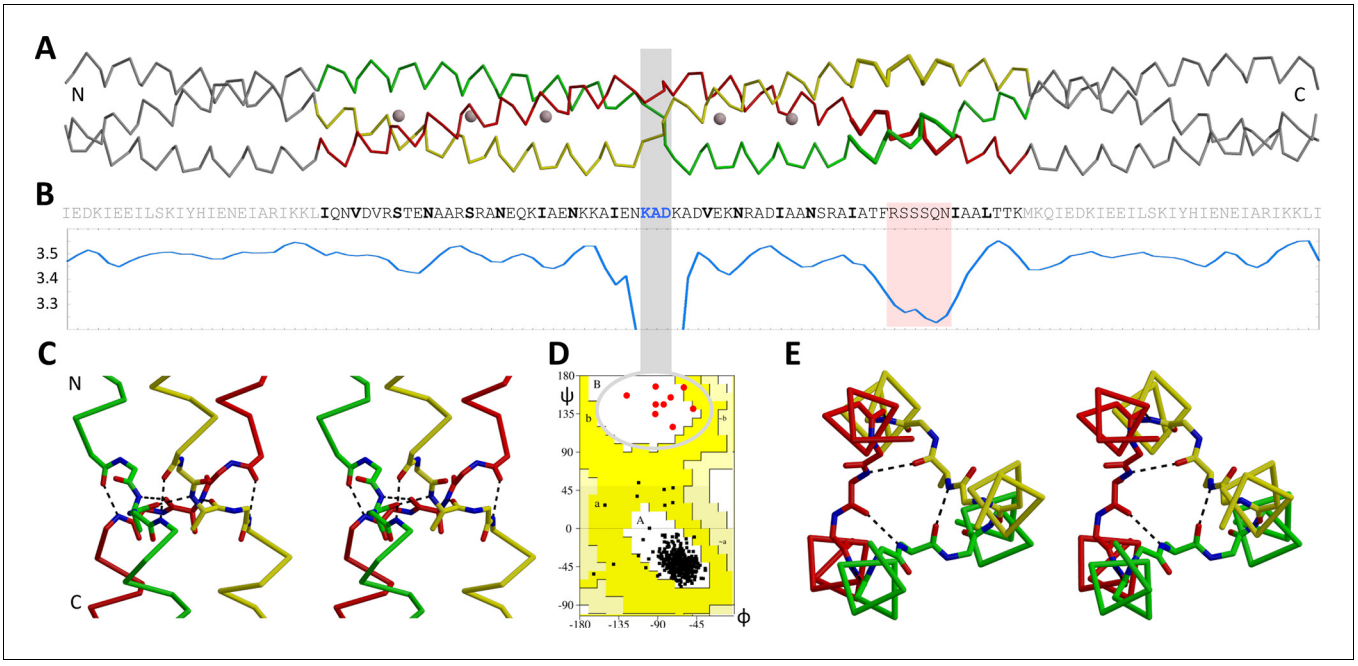


Figure 2. The β-layer in the Actinobacillus OMP100 stalk. (**A**) The structure of the Actinobacillus OMP100 stalk construct is aligned with (**B**) its sequence and a periodicity plot. The area of the stammer is highlighted in pink, the three residues of the β-layer by a grey bar. This bar points to the β region of the Ramachandran plot (**D**), where all nine β-layer residues of the trimer are found. The close-ups show the (**C**) side and (**E**) top view in stereo, highlighting the β-layer interactions. The trimer is colored by chain, GCN4 adaptors in grey. The plot is smoothed over a window of three residues to mask local fluctuations. Empty regions of the Ramachandran plot are cropped.

DOI: 10.7554/eLife.11861.005

perpendicular to the coiled-coil axis. Thereby only the central three residues of the motif, KAD, deviate from α-helical structure (***Figure 2***). All three fall into the β region of the Ramachandran plot, but only the central residue, alanine, forms backbone hydrogen bonds with the alanine residues of the other chains. We call this structural element a β-layer. It is essentially the same β-layer we described as an adaptor between α-helical and β-stranded segments of TAAs (***Hartmann et al., 2012***). Here it directly connects two α-helical segments, where the C-terminal one is rotated counterclockwise by ~120˚ around the trimer axis, as viewed from the N-terminus (***Figure 2***).

The first three residues of the IENKADKAD sequence motif occupy heptad positions *a, b* and *c* of the N-terminal α-helical segment, the last three residues positions *e, f* and *g* of the C-terminal segment. Therefore the β-layer, formed by the three central residues KAD, occurs in place of position *d*. The two segments are stabilized in their relative orientation by backbone hydrogen bonds from the last (*c* position) residue of each N-terminal helix to the first (*e* position) residue in the C-terminal helix of the neighboring chain (***Figure 2C***). This extends the continuous backbone hydrogen-bond network of each α-helix across the chains.

The nature of the discontinuity represented by this β-layer is related to the nature of stammers, but its effects are much stronger. With the insertion of 3 residues, stammers constitute a major strain on the conformation of the constituent helices of the coiled coil. In all examples to date, the resulting overwinding of the helices is absorbed by a short 3_{10}-helical segment. While these stammers can be best described to be part of a decad, the β-layer in OMP100 occurs in a motif of nine residues, a nonad. As the requirements of a nonad on its helices would be even more extreme than those of a decad, the strategy for its accommodation is a local but complete departure from helical structure.

β-layers in GCN4 fusions

Given the structural simplicity of β-layers, we wondered whether these could be brought about more generally by insertions of 2 residues into heptad coiled coils. Furthermore we wondered whether insertions of 6 residues, which pose similar demands on the coiled coil (***Figure 1***), also lead to the formation of β-layers. To tackle these questions experimentally, we designed a set of

constructs that had either 6 or 9 residues inserted between two consecutive GCN4 N16V adaptors, based on two different sequence motifs (***Figure 3***, ***Table 1***). One motif is IENKADKAD from *Actinobacillus* OMP100. The other, MATKDDIAN, is from a second family of prokaryotic coiled-coil proteins that we found to contain nonads and related periodicities; it occurs for example in 14 consecutive repeats in the protein Tcar0761 of *Thermosinus carboxydivorans*. From *Actinobacillus* OMP100 we derived the constructs A9 with the full IENKADKAD motif and A6 with the shortened motif IENKAD, as well as the 'control' construct A7 with the 7-residue motif IENKKAD. From *Thermosinus* Tcar0761 we derived the constructs T9, with the full MATKDDIAN motif, and T6, with the shortened motif MATKDD. The GCN4 N16V variant can form both dimeric and trimeric coiled coils and was chosen for these constructs to test for the oligomerization specificity of the inserts. All five constructs were resistant to proteolysis by proteinase K, showed typical α-helical CD spectra, did not melt upon heating to 95°C and yielded well-diffracting crystals. The structures of all constructs were trimeric and could be solved by molecular replacement, using the trimeric GCN4 structure as a search model. For T9, two structures were solved in alternative conformations (***Figure 3***). Apart from A7, which carries a heptad insert, all structures formed β-layers. These are identical in their structure (***Figure 4***), although they are not accommodated in the same way.

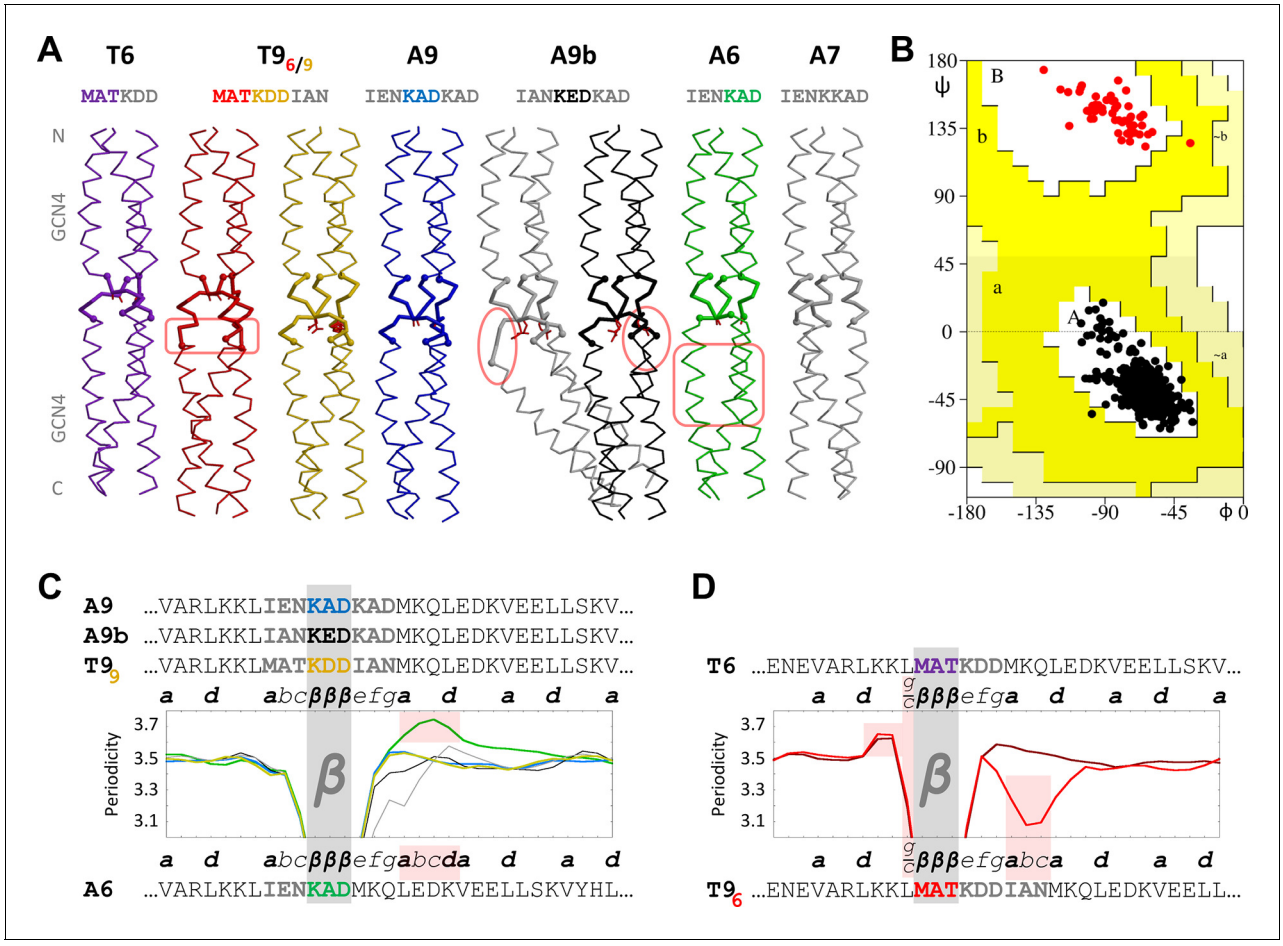


Figure 3. β-layers in 6- and 9-residue motifs between GCN4 adaptors. (**A**) The sequences and structures of the GCN4-fusion constructs are shown together with (**B**) a Ramachandran plot of their backbone torsion angles and (**C, D**) their periodicities. In the structures, the inserts between the GCN4 adaptors are drawn with thick lines. Disturbances in the α-helical segments are highlighted in pink; the stutter in the A6 structure and the stammer in the $T9_6$ structure are also highlighted in pink in panels C and D. In the periodicity plots, all proteins are aligned on the β-layer and their coiled-coil registers are indicated. The plots are shown separately for β-layers forming nonads (C) and hexads (D). A glitch in the periodicity caused by the *g/c* position preceding β-layers in hexads is highlighted in pink in panel D. As in the previous figure, the periodicity plots are smoothed over a three-residue sliding window. The Ramachandran plot in panel B includes all structures except the kinked grey A9b structure; all residues of the β-layers are shown as red dots and all other residues as black dots. Again, empty regions of the Ramachandran plot are cropped.

DOI: 10.7554/eLife.11861.006

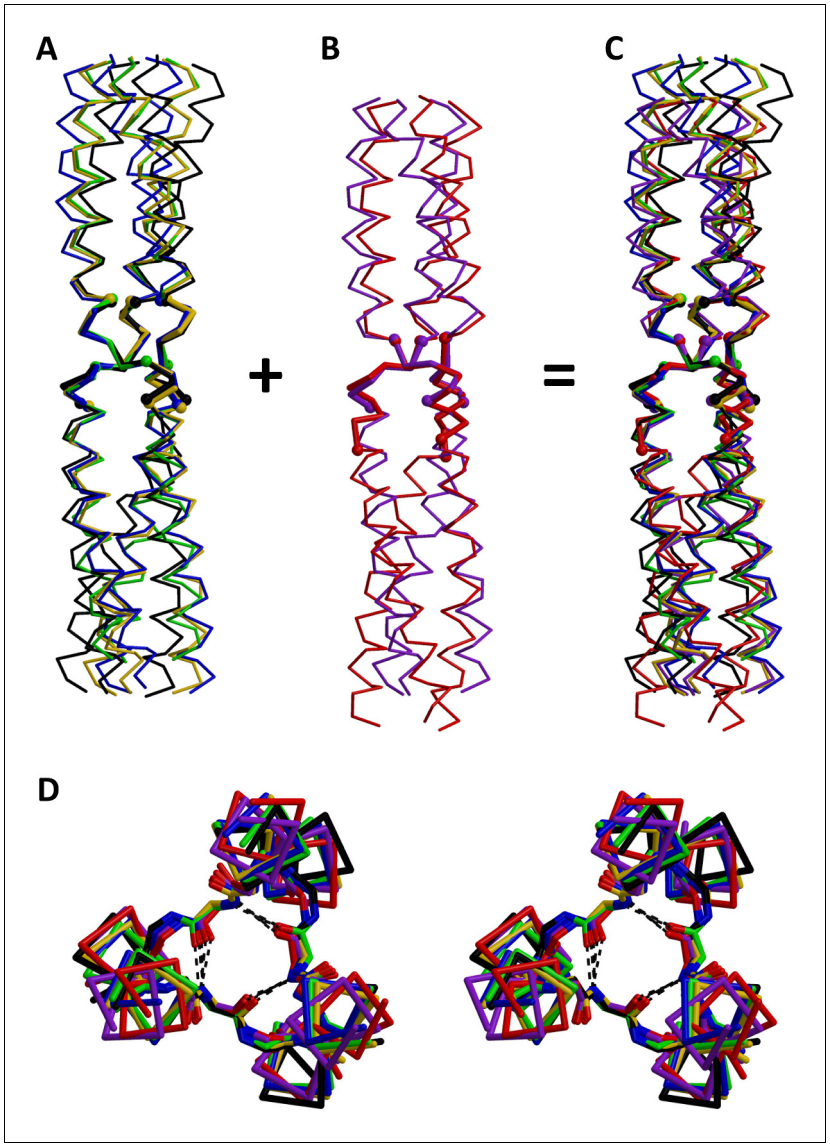


Figure 4. Superimposition of β-layers. All structures of β-layers between GCN4 adaptors were superimposed on the actual β-layer elements. Superimpositions are shown separately for β-layers occurring (**A**) in nonads ($T9_9$, A9, A9b, A6) and (**B**) in hexads (T6, $T9_6$); the kinked A9b structure (grey in ***Figure 3***) is omitted. Panel (**C**) shows all β-layers together. (**D**) Stereo view of the β-layer region in panel C, seen from the N-terminus. The structures are colored as in ***Figure 3***.

DOI: 10.7554/eLife.11861.007

In A9 and in one of the T9 structures, $T9_9$, the β-layers are formed as in *Actinobacillus* OMP100. The first three and last three residues of the insert are in heptad register with the flanking GCN4 adaptors and therefore constitute positions *a, b, c* and *e, f, g*. The middle three residues, KAD in A9, KDD in $T9_9$, form the β-layer in place of position *d* (***Figure 3C***). The A6 structure follows the same principle: again the first three residues of the insert are in heptad register with the N-terminal GCN4-adaptor, constituting positions *a, b* and *c*. The other three residues of the insert, KAD, form the β-layer in place of position *d*. As a consequence, the register of the C-terminal GCN4-adaptor is shifted at the junction to the insert, starting with position *e* instead of position *a*. This register conflict is resolved further downstream by the formation of a hendecad (highlighted in pink in ***Figure 3A and C***), so that the second half of the C-terminal adaptor retains its original register. In essence, the A6 structure shows a 9-residue element with the same structure as those found in A9

and $T9_9$, where the sequence IENKADMKQ borrows the last three residues from the C-terminal CGN4 adaptor and changes the periodicity of the latter at the junction.

In contrast, the T6 structure shows a 'real' 6-residue element. Here, the β-layer is formed by the first three residues of the insert, MAT (***Figure 3D***). The last three residues of the insert assume geometrically clear *e, f* and *g* positions and the C-terminal GCN4 adaptor follows in its native register, starting with an *a* position. Therefore, the β-layer occurs again in place of position *d*. A conflict with the native register of the N-terminal adaptor is avoided with just a small 'twist' to the adaptor's last residue: The C-terminal leucine, natively occurring in position *g*, is rotated outward from the core of the bundle by about 15° so that its Crick angle is biased towards the angle of a *c* position. In ***Figure 3*** this is noted as a *g/c* position, as it is close enough to an ideal position *g* for the preceding coiled coil to stay in register and close enough to a position *c* for the formation of the subsequent β-layer (highlighted in pink in ***Figure 3D***). Surprisingly, the alternative T9 structure, $T9_6$, starts out in the same way, forming the β-layer with the first three residues of the insert, MAT, after a *g/c* position. Consequently, the middle three residues constitute positions *e, f* and *g*. It thereby shows the same 6-residue element as the structure T6. The last three residues of the insert are accommodated as a sharply localized stammer, before the C-terminal adaptor starts in position *a* (highlighted in pink in ***Figure 3A and D***). Thus, the two structures T6 and $T9_6$ show that 6-residue elements are accommodated as N-terminally shortened 9-residue elements. While β-layers seem to strictly dictate the downstream register to start with an *e* position, they can occur after both *c* and *g* positions.

The observation that the T9 construct could accommodate the MATKDDIAN insert in two ways, forming the β-layer either at MAT or at KDD, led us to wonder whether the same could happen with the A9 insert, IENKADKAD, if the glutamate was interchanged with the central alanine to mirror the first six residues of the T9 insert (A9b, IANKEDKAD). We had previously found that β-layers which occur as connectors between TAA domains prefer small, hydrophobic residues in their central position (***Hartmann et al., 2012***; ***Bassler et al., 2015***). We therefore thought that the central aspartate of the T9 insert might have been sufficiently unfavorable ($T9_9$) that an alternative, with the alanine of MAT at the center of the β-layer ($T9_6$), became observable, even though $T9_9$ allows the flanking coiled-coil segments to remain unperturbed and $T9_6$ requires their distortion. We reasoned that the larger glutamate residue at the center of A9b might even be sufficiently unfavorable to move this construct quantitatively to the alternative structure, with the β-layer formed over the first three residues (IAN). A9b in fact crystallized in two alternative structures (***Figure 3***), but in both the β-layer formed over the central glutamate. Since this residue was indeed too large and polar to be accommodated without distortion, the first turns of the downstream helices are perturbed to different extents in both instances, leading to a pronounced kink in one of the structures (highlighted in pink in ***Figure 3A***). We were surprised to see that the penalty introduced by the central glutamate was not sufficient to produce the alternative structure; the reasons for this are unclear to us at present.

The α/β coiled coil

With the expectation to obtain a continuous fiber of alternating α and β elements, we built a construct with repeating nonads, based on *Thermosinus* Tcar0761 (***Figure 5***). The 14 consecutive, almost perfect MATKDDIAN repeats in this protein are flanked by long heptad segments. In our construct we omitted the middle ten nonad repeats and trimmed the N- and C terminal heptad segments for in-register fusion to GCN4-N16V (***Table 1***; red sequence in ***Figure 5***). Crystallization trials yielded crystals in space group $P6_3$, diffracting to a resolution of 1.6Å, with one chain in the asymmetric unit and the trimer built by crystallographic symmetry around the c axis. The structure could be solved by molecular replacement using fragments of the T6 and A9 structures. It shows the anticipated α/β coiled coil with four consecutive β-layers. These layers are formed by the residues MAT of the repeats; the other residues, corresponding to KDDIAN, constitute positions *e,f,g,a,b,c* of the segments between the β-layers. Therefore, in accordance with heptad notation, the repeats can be written as IANMATKDD, with the isoleucine forming classical hydrophobic *a* layers and the MAT forming β-layers in place of position *d* (***Figure 5***). Only the first β-layer is part of a 6-residue element (hexad) and occurs after a position *g* of the preceding heptad. This *g* position is biased towards a *c* position, as described above for the structures T6 and $T9_6$, yielding the same *g/c* position. With its alternating *a*- and β-layers, the α/β coiled coil is a new class of protein fiber, based on a novel supersecondary structure element.

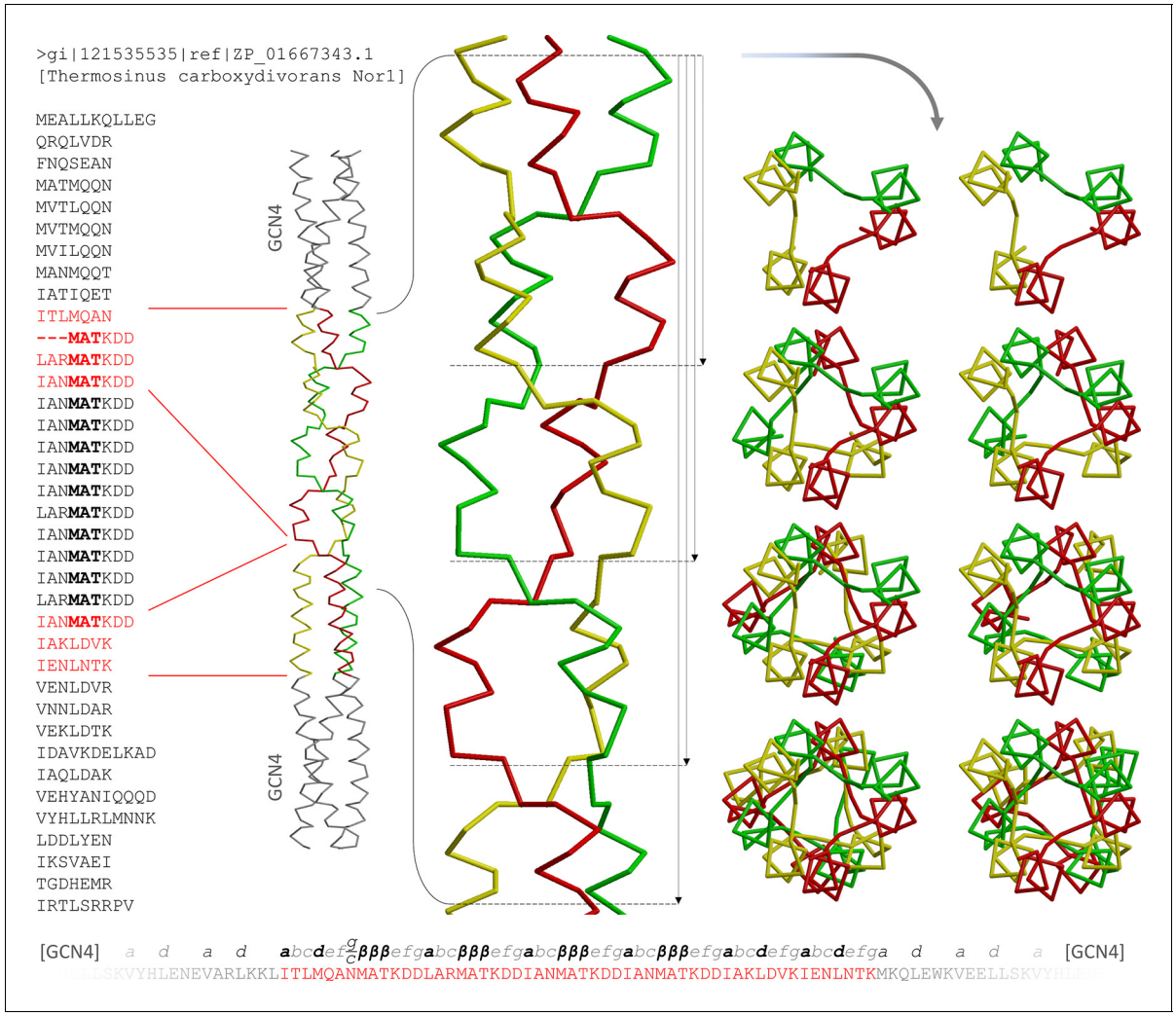


Figure 5. The α/β coiled coil in the Tcar0761 construct. The two regions fused between GCN4 adaptors in our construct are shown in red on the full sequence of Tcar0761 (left). Next to the sequence, the structure is depicted as a Cα-trace and the four consecutive β-layers are enlarged. On the right, top views are shown, looking down the bundle from the N-terminus. As indicated by the arrows next to the side view, they show 1, 2, 3 or all 4 β-layers. At the bottom, the sequence of the construct is shown together with the assigned register.

DOI: 10.7554/eLife.11861.008

The α/β coiled coil of *Thermosinus* Tcar0761 is built of nonads and thus contains six residues per repeat in the α region of the Ramachandran plot, which retain one backbone hydrogen bond characteristic of α-helical structure. We think that it should be possible to reduce this structure by removing three α-helical residues and thus the single remaining backbone hydrogen bond from the parent structure. Such a minimalistic α/β coiled coil would be built of hexads, with three residues in the β and three in the α region of the Ramachandran plot. We have not so far detected coiled-coil proteins with β-layers in hexad spacing, nor have we been successful in constructing such a structure by fusion of MATKDD repeats between GCN4-N16V adaptors. However, as we will show in the next section, a tail-fiber protein from a *Streptococcus pyogenes* prophage (2C3F) contains an α/β coiled coil with four β-layers, two of which are in a hexad spacing.

β-Layers in proteins of known structure

At the beginning of this project we had identified nonads in the stalks of TAAs and in the N-terminal coiled coils of a family of prokaryotic endonucleases listed in Pfam as PD-(D/E)XK, specifically in the crenarchaeal representatives of this family. The bacterial representatives, where they had the coiled-coil stalk, lacked nonads or related periodicities (in Pfam however, all the coiled-coil segments of

this family are grouped together in entry DUF3782). Surprisingly, we found that some bacteria contain coiled-coil proteins that lack the endonuclease domain, but are very similar to the coiled coils of the crenarchaeal proteins; *Thermosinus* Tcar0761 belongs to these. The β-layers in this family have the consensus sequence [aliphatic]–A–T–K–[polar]–[DE] (***Figure 6***). Pattern searches with this motif

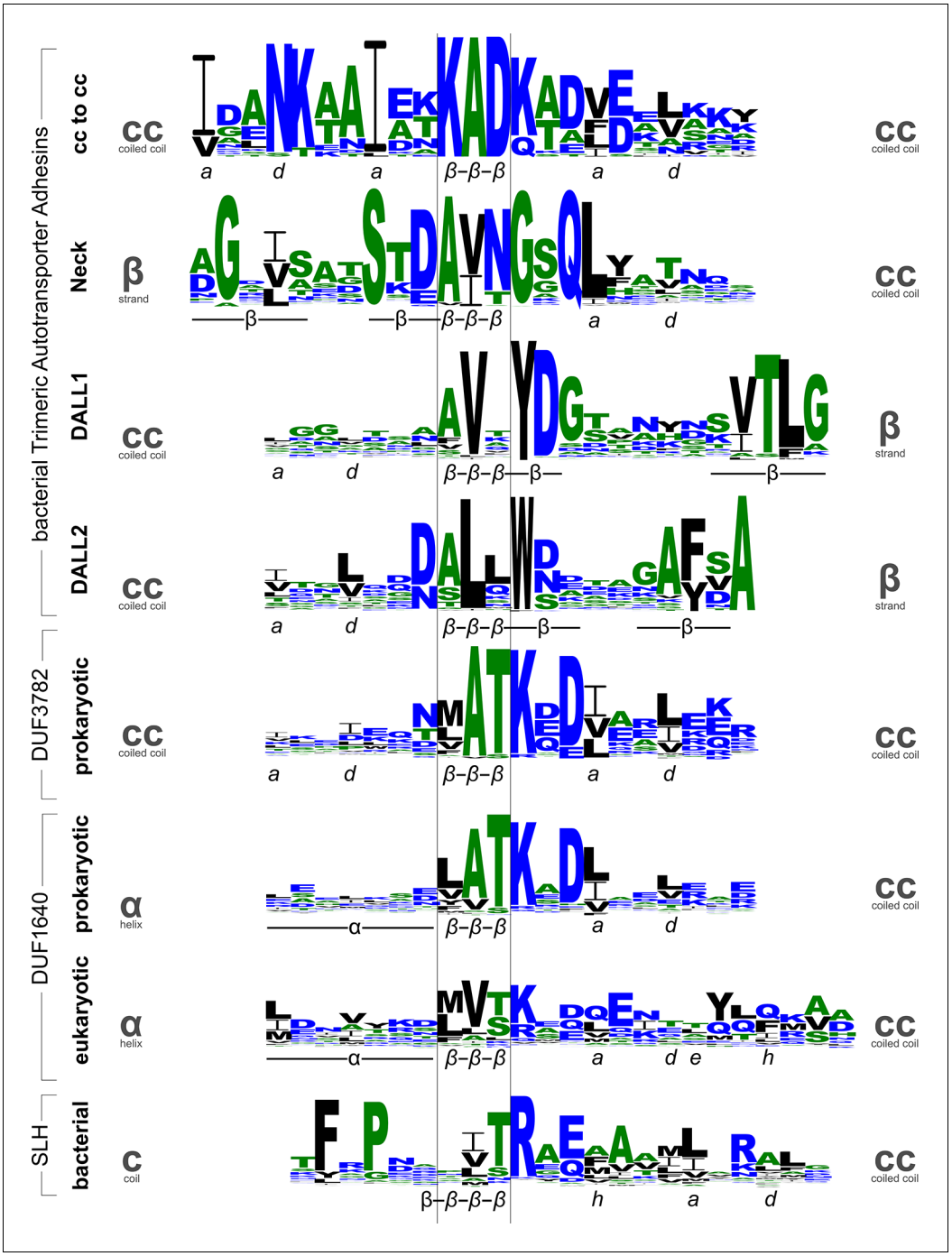


Figure 6. Sequence logos for β-layers in different protein families. The sequence logos show the conservation patterns of β-layers and their adjacent secondary structure elements in domains of Trimeric Autotransporter Adhesins (stalk, neck, and two variants of the DALL domain), the DUF3782 family of prokaryotic endonucleases, the DUF1640 family of membrane proteins from prokaryotes and organelles, and the surface layer homology (SLH) domain of bacteria. Annotations of the secondary structure (α: helix, β: strand) and coiled-coil register are shown beneath the logos. Grey symbols on the sides indicate the type of secondary structure transition mediated by the β-layer.

DOI: 10.7554/eLife.11861.009

led us to the discovery of a family of integral membrane proteins found in prokaryotes and mitochondria (DUF1640), which carry this motif prominently at the beginning of their C-terminal stalk (***Figure 6***). However, our sequence searches, both based on sequence patterns and on the discovery of relevant insertions into the heptads of coiled coils, have progressed slowly, as they require much case-by-case analysis. This is due on the one hand to the frequency of the β-layer sequence patterns in coiled coils and on the other to the difficulty of establishing reliably the local register of coiled coils that deviate from the heptad repeat with existing software. Indeed, as we have described for TAAs (***Szczesny and Lupas, 2008***; ***Bassler et al., 2015***), many of these escape detection entirely with current programs.

Given the structural identity between the β-layers resulting from hexads and nonads in coiled coils, and the supersecondary structures we characterized at the transition between coiled-coil segments and β-stranded domains in TAAs (***Hartmann et al., 2012***), we searched systematically for other instances of β-layers in proteins of known structure. The results are collected in ***Figure 7*** and ***Table 2***. All proteins we identified are homotrimers, except for the SLH domain, which is a monomer with pseudo-threefold symmetry; a majority are from viruses, mainly bacteriophage. Most β-layers occur in the context of coiled coils and we have termed these 'canonical'. They are usually found capping one of the ends of the coiled coil, more often the N-terminal than the C-terminal one, and we have found only two further examples of coiled coils with internal β-layers: MPN010, a protein of unknown function from *Mycoplasma pneumoniae* (2BA2), and the aforementioned tail-fiber protein with hyaluronidase activity from the *Streptococcus pyogenes* prophage SF370.1 (2C3F). The latter contains a coiled coil with four β-layers, one near the N-terminus, two internal, and one at the C-terminal end; the two internal β-layers have the sequence LQQ**KAD**KET**VYT**KAE and are thus in a hexad spacing, with the first resembling the β-layer sequence of OMP100 and the second the one of Tcar0761. Remarkably, this second β-layer deviates from canonical β-layer structure, which we attribute to the serine in the first core position of the downstream coiled coil. This serine spans a water network in the core of the trimer, which invades the β-interactions of the β-layer with bridging waters, leading to a largely increased diameter of the layer. This wider diameter might be further promoted by the bulky tyrosine side chain of the central β-layer residue, which is bent out of the core. Nevertheless, such tandems with the consensus sequence Lxx**KAD**Kxx**VYT**KxE occur in many bacterial ORFs (also in some TAAs, such as *Neisseria meningitidis* NadA4) and thus probably constitute a co-optimized module.

A structural analysis of canonical β-layers in light of their conserved sequence patterns (***Figure 6***, ***Table 2***) shows that they favor hydrophobic residues in β_1 and β_2 (***Figure 6***), and particularly the β_2 residue tends to be of smaller size (i.e. A or V). They can follow upon either position *a* or *d* of the preceding coiled coil, but always lead into positions *e, f, g* of the following coiled coil. Thus, when they follow upon position *a* they yield the register *a-b-c-*β_1*-*β_2*-*β_3*-e-f-g* (seen in nonads), whereas when they follow upon position *d* they bias the residue in position *g* towards *c* to yield the register *e-f-g/c-*β_1*-*β_2*-*β_3*-e-f-g* (seen in hexads). For the purpose of the following discussion we will refer to these two registers collectively as α_1-α_2-α_3-β_1-β_2-β_3-e-f-g.

For β-layers that occur at the C-terminal end of coiled coils (for example in the DALL1 and DALL2 domains of TAAs), the flanking residues do not form conserved mainchain or sidechain interactions with the layer or with each other, and their conservation pattern is dominated by interactions with the downstream domain. Since β-layers can form interaction networks that provide a C-cap to the preceding coiled coil (see below), it is surprising that they do not do so in most structures where they occur at the C-terminal end of coiled coils.

For β-layers that occur at the N-terminal end (for example in the necks of TAAs or in influenza hemagglutinin HA_2), the β_3 residue acts as an N-cap for the following helix, coordinating the backbone NH group of residue *g* (***Figure 8***); it is thus almost always D, N, T, or S (the capping role of this residue has been described in detail in the fusion-pH structure of influenza hemagglutinin HA_2 (***Chen et al., 1999***). In return, the sidechain of the residue in position *g* forms a hydrogen bond with the backbone NH group of the β_3 residue, closing a ring of sidechain-backbone interactions between these two residues; it is thus almost always D, E, or Q. Where it is D or E, it can further form a salt bridge to the residue in position *e* of the neighboring chain (clockwise as viewed from the N-terminus), which is broadly conserved as K or R. This residue essentially always forms either this salt bridge, or a hydrogen bond with the backbone carbonyl group of the β_1 residue, as depicted in ***Figure 8***. This interaction network allows β-layers to form stably at the N-terminal end of

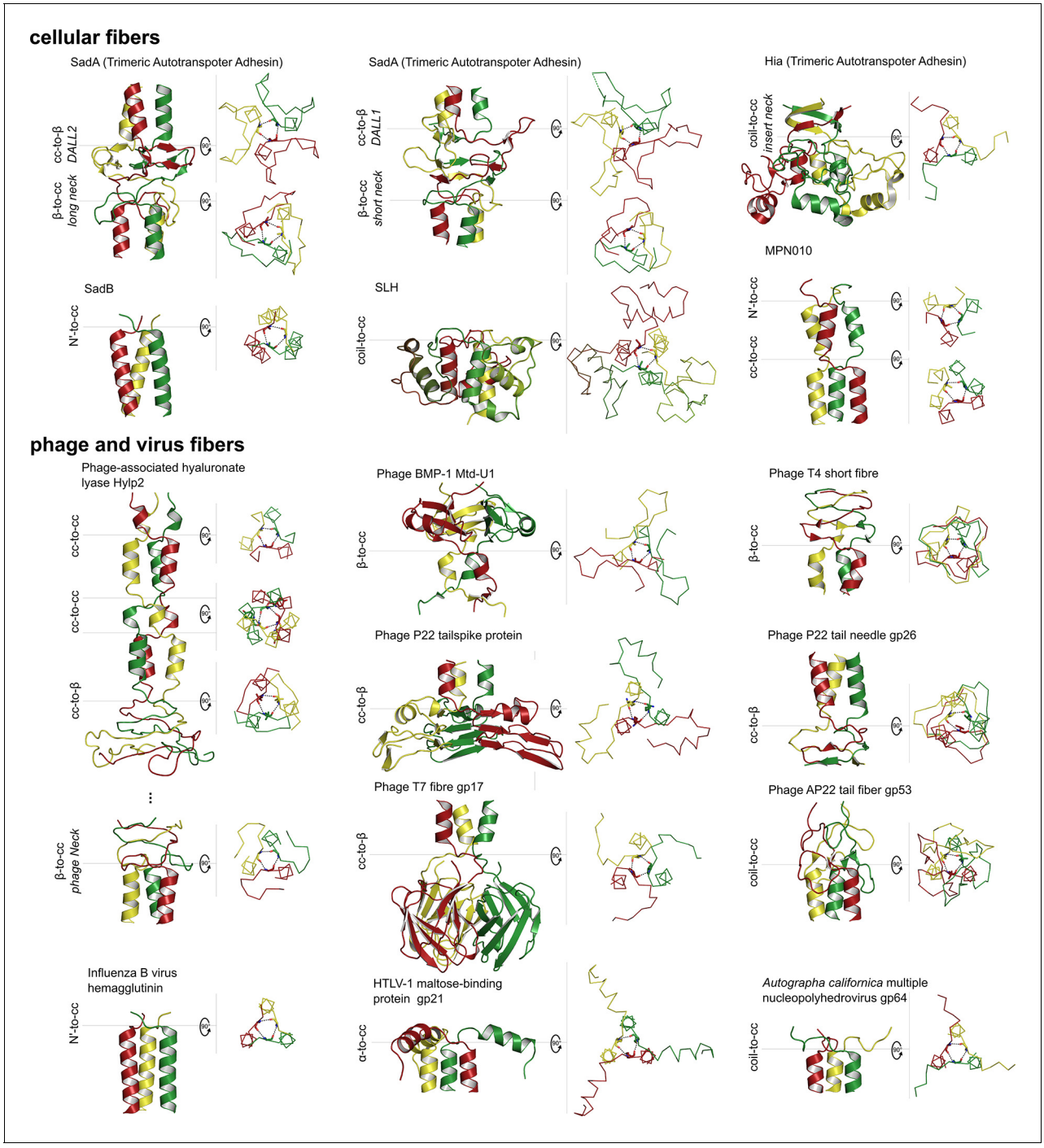


Figure 7. Gallery of canonical β-layers in proteins of known structure. The parts of the structures containing β-layers are shown in side view (cartoon depiction, left) and the β-layers in top view (backbone trace, right), with their central (β_2) residues in stick representation. ***Table 2*** lists the detailed information for the presented proteins.

DOI: 10.7554/eLife.11861.010

coiled-coil proteins, as seen in the crystal structures of SadB, MPN010, and the fusion-pH structure of influenza hemagglutinin HA_2.

β-Layers that occur within coiled coils show substantially the same interactions and conservation patterns as the ones that act as N-caps, when the residue in β_1 is hydrophobic. Only occasionally, the K or R in position *e* shows yet a third conformation, coordinating the backbone carbonyl of the α_2 residue of the preceding helix in the neighboring chain (counterclockwise), thus providing a C-

Table 2. *β*-Layers in proteins of known structure.

Cellular proteins (canonical)

PDB	Type	Protein	Domain	Species	Sequence	Similar structures
2YO3	cc-to-β	SadA	TAA DALL1	*Salmonella enterica*	***a***bc***d***efg***βββ***EEEECC 1306-LKASEAG**SVR**YETNAD-1321	3WPA, 3WPO, 3WPP, 3WQA (*Acinetobacter sp. Tol5*), 4USX (*Burkholderia pseudomallei*)
2YO2	cc-to-β	SadA	TAA DALL2	*Salmonella enterica*	***a***bc***d***efg***βββ***EECCC 310-VAGLAED**ALL**WDESI-324	3ZMF, 2YNZ (*Salmonella enterica*)
2YO3	β-to-cc	SadA	TAA Short neck	*Salmonella enterica*	EEEECCC***βββ***efg***a***bc***d***efg***h***ijk***l***mno 1345-AAVNDTD**AVN**YAQLKRSVEEANTYTDQK-1372	4LGO (*Bartonella quintana*), 3WP8, 3WPA, 3WPR (*Acinetobacter* sp. Tol5), 1P9H (*Yersinia enterocolitica*), 2XQH (*Escherichia coli*), 3D9X (*Bartonella henselae*), 2YO0 (*Salmonella enterica*), 3S6L, 4USX (*Burkholderia pseudomallei*), 2GR7 (*Haemophilus influenzae*)
2YO2	β-to-cc	SadA	TAA Long neck	*Salmonella enterica*	EEEE***βββ***efg***a***bc***d***efg 349-DSTD**AVN**GSQMKQIEDK-365	2YNZ, 3ZMF (*Salmonella enterica*), 3EMO (*Haemophilus influenzae*), 3LAA, 3LA9, 4USX (*Burkholderia pseudomallei*), 3WPA, 3WPO, 3WPP, 3WPR, 3WQA (*Acinetobacter* sp. Tol5), 3NTN, 3PR7 (*Moraxella catarrhalis*)
1S7M	β-to-cc	Hia	TAA Insert neck 1	*Haemophilus influenzae*	EE***βββ***efg***a***bc 642-NT**AAT**VGDLRG-652	3EMF (*Haemophilus influenzae*)
4C47	Nterm-to-cc	SadB	-	*Salmonella enterica*	CC***βββ***efg***a***bc***d***efg 23-DY**FAD**KHLVEEMKEQ-37	-
5APP	cc-to-cc	OMP100	TAA Stalk	*Actinobacillus actinomycetemcomitans*	***a***bc***d***efg***a***bc***βββ***efg***a***bc***d***efg 153-IAENKKAIEN**KAD**KADVEKNRAD-175	-
5APZ	cc-to-cc	Tcar0761	DUF3782	*Thermosinus carboxydivorans*	***a***bc***d***efg***βββ***efg***a***bc 68-ITLMQAN**MAT**KDDLAR-83	-
2BA2	Nterm-to-cc	MPN010	DUF16	*Mycoplasma pneumoniae*	CCC***βββ***efg***h***ijk 5-GTR**YVT**HKQLDEK-17	-
2BA2	cc-to-cc	MPN010	DUF16	*Mycoplasma pneumoniae*	***h***ijk***a***bc***βββ***efg***a***bc***d***efg***h***ijk 14-LDEKLKN**FVT**KTEFKEFQTVVMES-37	-
3PYW	coil-to-cc	S-layer protein Sap	SLH	*Bacillus anthracis*	CCCCE***βββ***efg***h***ijk***a***bc***d***ef 35-FEPGK**ELT**RAEAATMMAQILN-55 ... 94-FEPNG**KID**RVSMASLLVEAYK-114 ... 156-WEPKK**TVT**KAEAAQFIAKTDK-176	-
Phage and virus proteins (canonical)						
2C3F	cc-to-cc	Tail fiber hyaluronidase	-	*Streptococcus pyogenes* (prophage SF370.1)	***a***bc***βββ***efg***h***ijk***a***bc***βββ***efg***βββ***efg***h***ijk 69-IDG**LAT**KVETAQKLQQ**KAD**KET**VYT**KAESKQE-99	2DP5 (*Streptococcus pyogenes*)
2C3F	cc-to-β	Tail fiber hyaluronidase	-	*Streptococcus pyogenes* (prophage SF370.1)	***d***efg***a***bc***βββ***CEEEEE 97-SKQELDK**KLN**LKGGVM-112	2DP5 (*Streptococcus pyogenes*)

Table 2 continued on next page

Table 2 continued

Cellular proteins (canonical)

PDB	Type	Protein	Domain	Species	Sequence	Similar structures
2C3F	β-to-cc	Tail fiber hyaluronidase	TAA short neck homolog	*Streptococcus pyogenes* (prophage SF370.1)	EEEECCE***βββ***efg***h***ijk***a***bc***d***efg 310-DPTANDH**AAT**KAYVDKAISELKKL-327	2DP5, 2WH7, 2WB3 (*Streptococcus pyogenes*)
4MTM	coil-to-cc	gp53	-	Bacteriophage AP22	CCCCE***βββ***efg***a***bc***d***efg 155-NDVGS**ALS**AAQGKVLNDK-172	-
1YU4	β-to-cc	Major tropism determinant U1 variant (Mtd-U1)	-	*Bordetella* Phage BMP-1	CCCCEE***βββ***efg***a***b 41-TAGGFP**LAR**HDLVK-54	-
1TSP	cc-to-β	Tailspike protein	Phage P22-tail	Phage P22	***de***fg***h***ijk***βββ***EEE 113-YSIEADKK**FKY**SVK-126	1CLW, 2XC1, 2VFM, 2VFP, 2VFQ, 2VFO, 2VFN [...] (Phage P22) 4OJP, 4OJ5, 4OJL [...] (*E. coli* Bacteriophage CBA120) 2V5I (Bacteriophage Det7), 2X3H (Enterobacteria phage K1-5)
2POH	cc-to-β	Phage P22 tail needle gp26	-	Phage P22	***a***bc***d***efg***βββ***CEEC 133-ISALQAD**YVS**KTAT-146	3C9I, 4LIN, 4ZKP, 4ZKU, 5BU5, 5BU8, 5BVZ (Phage P22)
1H6W	β-to-cc	Short fiber	Receptor binding domain	Bacteriophage T4	EEEEEECCEE***βββ***efg***a***bc***d***e 321-MTGGYIQGKR**VVT**QNEIDRTI-341	1OCY, 1PDI, 2XGF, 2FKK, 2FL8 (Bacteriophage T4)
4A0T	cc-to-coil	gp17	gp37_C	Bacteriophage T7	c***de***fg***h***ijk***βββ***CCCC 454-WLDAYLRDS**FVA**KSKA-469	4A0U (Bacteriophage T7)
1MG1	α-to-cc	Maltose-binding protein GP21	TLV_coat	Primate T-lymphotrophic virus 1 (HTLV-1)	HHHHHHE***βββ***efg***a***bc***de***fg***h***ijk 364-AAQTNAA**AMS**LASGKSLLHEVDKD-387	-
3DUZ	coil-to-cc	GP64	Baculo_gp64	*Autographa californica* Multiple Nucleopolyhedrovirus	CCC***βββ***efg***a***bc***d***efg 293-EGD**TAT**KGDLMHIQEE-308	-
4NKJ	Nterm-to-cc	Hemagglutinin	Hemagglutinin HA_2	Influenza B virus	E***βββ***efg***a***bc***de***fg***h***ijk 4-V**AAD**LKSTQEAINKITKN-21	1QU1 (Influenza A virus)
Unusual β-layer proteins						
4NQJ	α-to-α	TRIM Ubiquitin E3 ligase	DUF3583	*Homo sapiens*	HHHHHHH***βββ***HHHHHHH 143-SVGQSKE**FLQ**ISDAVHF-159	-
2F0C	(cc-to-) coil-to-β	Receptor binding protein (ORF49)	-	*Lactophage* tp901-1	***a***bc***d***efg***a***bCCCC***βββ***CEEC 22-LEAINSELTSGGN**VVH**KTGD-41	3D8M, 3DA0 (Lactophage tp901-1)
1AA0	(cc-to-) coil-to-β	Fibritin	Fibritin_C	Bacteriophage T4	***a***bc***d***efgC***βββ***EEEEE 450-VQALQEAG**YIP**EAPRD-465	1AVY, 2BSG, 2IBL, 2WW6, 2WW7, 3ALM (Bacteriophage T4), 5C0R (Influenza A), 2LP7 (Human Immunodeficiency Virus 1), 1NAY
2XGF	coil-to-coil	Long tail fiber needle	-	Bacteriophage T4	EEEECCCCCCCC***βββ***CCCCEEEE 934-EAWNGTGVGGNK**MSS**YAISYRAG-956	-

Table 2 continued on next page

Table 2 continued

Cellular proteins (canonical)

PDB	Type	Protein	Domain	Species	Sequence	Similar structures
1H6W	coil-to-coil-(to-β)	Short fiber	-	Bacteriophage T4	CCC***βββ***CCCCEEEEE 284-NAD**VIH**QRGGQTING-298	-
4UXG	β-to-coil	Proximal long tail fibre protein gp34	-	Bacteriophage T4	EEE***βββ***CCCCCC 1233-FVQ**VFD**GGNPPQ-1244	-
4UXG	α-to-coil	Proximal long tail fibre protein gp34	-	Bacteriophage T4	HHHHC***βββ***CCCEEE 1245-PSDIG**ALP**SDNATM-1258	-
3QC7	α-to-coil	Head fiber	-	Bacteriophage Phi29	HHHHHHH***βββ***CCCCCCC 221-NLRTMIG**AGV**PYSLPAA-237	-

DOI: 10.7554/eLife.11861.011

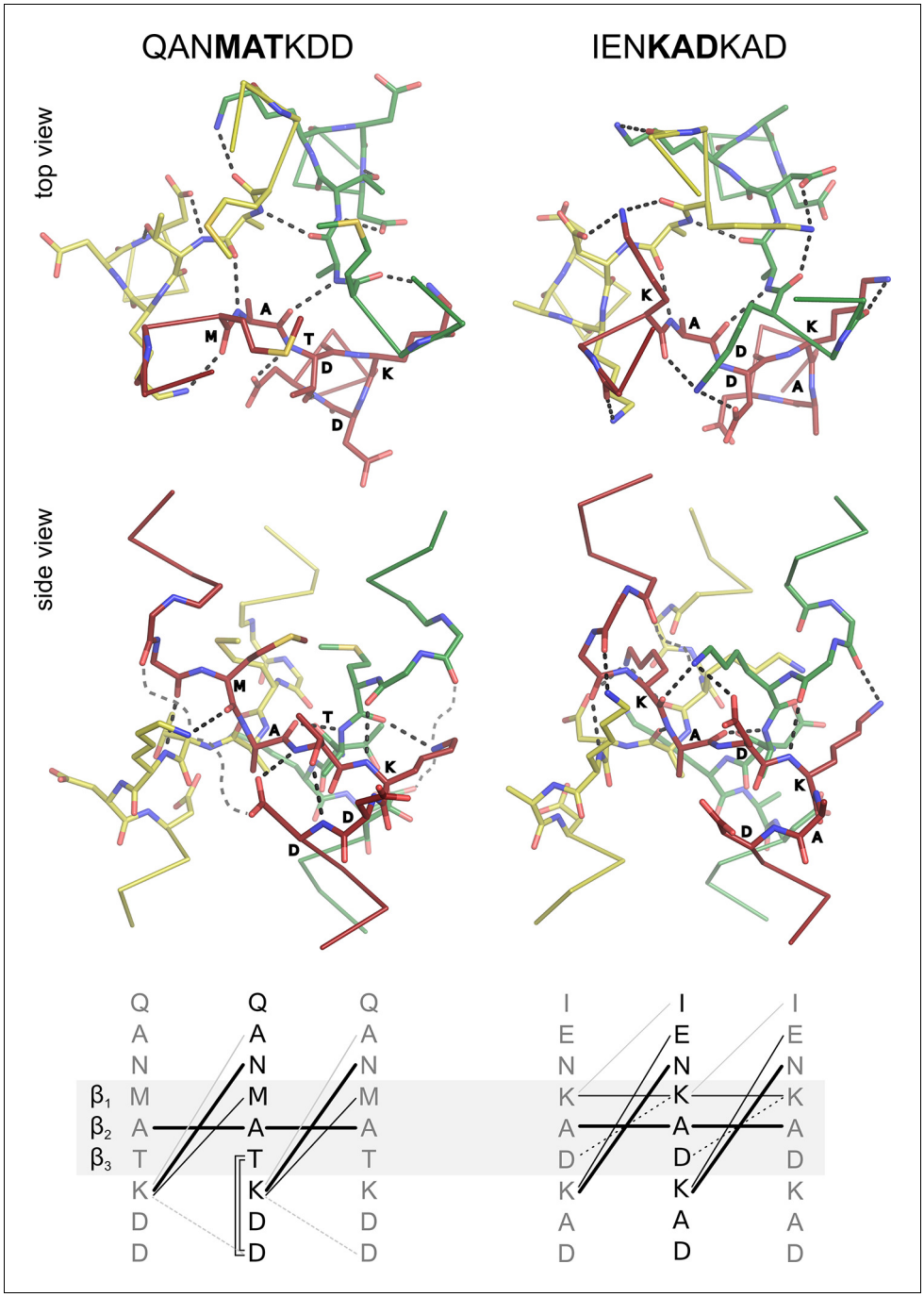


Figure 8. Interaction networks of canonical β-layers. The two distinct interaction networks of N-capping β-layers based on the sequence MATKDD and C-capping β-layers based on the sequence KADKAD are compared. The upper panels show the interactions at the first β-layer in the Tcar0761 structure and the β-layer in the A9 structure, both in top and side view. For clarity, the side views show only the interactions of the red chain. The lower panels show a schematic representation of the interactions: invariant backbone-to-backbone hydrogen bonds are drawn as bold lines, network-specific backbone-to-sidechain and sidechain-to-sidechain interactions are drawn as solid and broken lines, respectively. Grey lines indicate alternative/additional interactions which are not formed in the depicted β-layers but can be found in other instances as described in the main text. These interactions are indicated by loose grey broken lines in the side views.

DOI: 10.7554/eLife.11861.012

capping interaction. However, when the β_1 residue is K (mainly in the stalks of TAAs and related phage proteins) the interaction network changes entirely from an N-cap of the following helices to a C-cap of the preceding ones. The K in β_1 reaches across the core of the trimer to form one, two, or all three of the following interactions: coordinate the backbone carbonyls of the α_1 and β_1 residues, and the sidechain of the β_3 residue, all from the neighboring chain (clockwise). Additionally, the K or R in position *e* is entirely found in the C-capping conformation. In all cases, the network is completed by backbone hydrogen bonds from the $\alpha_3$residues to the residues in *e* (clockwise), as already described for OMP100 (***Figure 2C***). These considerations suggest that in a tandem of β-layers with hexad spacing, the first layer should favor a C-capping network, with K in position β_1, and the second an N-capping network, with a hydrophobic residue in β_1. This is in fact observed in the *Streptococcus* prophage tail-fiber protein (2C3F).

Conclusions

The range of periodicities that α-helical coiled coils can assume is limited by the strain they impose on the constituent helices, as they progressively deviate from the 3.63 residues per turn of an undistorted α-helix. Insertions of 3 residues into a heptad background (stammers, 10/3 = 3.33) lead to the largest strain observed so far in continuous coiled coils and are accommodated by the local distortion of the α-helix into a 3_{10} helix. We find that increasing the strain further by insertions of 2x3 or 3x3 residues leads to a complete loss of helical structure and the local formation of short β-strands. These cross to form a triangular plane, which moves the path of each chain by 120˚ counterclockwise around the trimer axis. Within this plane, the central residues of the three β-strands form backbone hydrogen bonds whose geometry deviates substantially from that seen in β-sheets. We have named them β-layers and show that they can be brought about in a straightforward way by the insertion of 6 or 2 (9 = 2 modulo 7) residues into a heptad background. We propose that β-layers offer two clear advantages to protein fibers. They increase their resilience by tightly interleaving the monomers within the fiber and they offer a simple mechanism to integrate β-stranded domains into these fibers, thus increasing their functional complexity. Our results show that a novel backbone structure is accessible to the 20 proteinogenic amino acids in the allowed regions of the Ramachandran plot with only minor mutations to a known fold.

Materials and methods

Cloning

If not otherwise indicated, constructs were amplified by primer extension. Primers used for amplification, cloning and mutagenesis are listed in ***Table 3***.

The OMP100 construct encompasses residues 133–198 of OMP100 from *Actinobacillus actinomycetemcomitans* (Genbank BAB86905.1), fused at both the N- and the C-terminus in heptad register to the trimeric leucine zipper GCN4-pII. The amplified DNA fragment was cloned in Eco31I-sites of pIBA-GCN4tri-His (***Hernandez Alvarez et al., 2008***).

The Tcar0761 construct is derived from open reading frame 0761 of *Thermosinus carboxydivorans* Nor1 (Genbank ZP_01667343.1). A DNA fragment encoding residues 68–101, fused directly to a fragment encoding 191–211, was made by gene synthesis (GenScript) and cloned in the Eco31I-sites of pIBA-GCN4 N16V-His.

The GCN4 N16V version of the pIBA-GCN4 series allows for the expression of protein fragments fused at both termini to GCN4 adaptors carrying the N16V mutation, a variant of the leucine zipper that forms a mixture of dimers and trimers. pIBA-GCN4 N16V-His was constructed by replacing the XbaI/HindIII fragment of pASK IBA2 by a DNA fragment containing the XbaI site, ribosomal binding site, N-terminal GCN4 N16V adaptor, multiple cloning site, C-terminal GCN4 N16V adaptor, (His)6-tag and the HindIII site. Aspartate residues in position *f* of the first heptad were replaced by methionine and tryptophan in the N- and C-terminal GCN4 adaptor as described before (***Deiss et al., 2014***).

Constructs A6, T6 and T9 were amplified and cloned into the NdeI and BamHI sites of the expression vector pETHis1a_Nde1, a modified version of pETHis1a (***Bogomolovas et al., 2009***) allowing for expression of the constructs with a C-terminal (His)6-tag and a TEV-protease cleavage site. A7

Table 3. Primers used in this study.

Construct	Primer
OMP100	**P omp1**: 5′-GACCATGGTCTCCGATTCAGAACGTGGATGTGCGCAGCACCGAAAACGCGGCGCGCAGCCGCGCGAACGAACAG **P omp2**: 5′-GCTTTATCCGCTTTGTTTTCAATCGCTTTTTTGTTTTCCGCAATTTTCTGTTCGTTCGCGCGGCTGC **P omp3**: 5′-GAAAACAAAGCGGATAAAGCGGATGTGGAAAAAAACCGCGCGGATATTGCGGCGAACAGCCGCGCGATTGCGACCTTTCG **P omp4**: 5′-GACCATGGTCTCCTCATTTTGGTGGTCAGCGCCGCAATGTTCTGGCTGCTGCTGCGAAAGGTCGCAATCGCGCG
pASK IBA GCN4 N16V	**P iba1**: 5′-ACAAAAATCTAGATAACGAGGGCAAAAAATGAAACAGCTGGAAATGAAAGTTGAAGAACTGCTGTCCAAAGTCTACCACCTGGAAAACGA **P iba2**: 5′-CTCGAGGGATCCCCGGGTACCGAGCTCGAATTCGGGACCATGGTCTCCCAGTTTTTTCAGACGCGCAACTTCGTTTTCCAGGTGGTAGAC **P iba3**: 5′-GTACCCGGGGATCCCTCGAGAGGGGGACCATGGTCTCAATGAAACAGCTGGAATGGAAAGTTGAAGAACTGCTGTCCAAAGTCTACCACC **P iba4**: 5′-CACAGGTCAAGCTTATTAGTGATGGTGATGGTGATGGCCAGAACCAACCAGTTTTTTCAGACGCGCAACTTCGTTTTCCAGGTGGTAGACTTTGGACAGC
T6	**T6 p1**: 5′-GGAATTCCATATGAAGCAGCTGGAAGACAAGGTGGAGGAACTGTGTCCAAAGTGTACCATCTGGAAAACGAGGTGGCGCGTCTGAAGAAG **T6 p2**: 5′-CTTGGACAGCAGTTCTTCCACCTTATCTTCCAGCTGCTTCAATCATCTTTGGTCGCCATCAGCTTCTTCAGACGCGCCACCTC **T6 p3**: 5′-GGTGGAAGAACTGCTGTCCAAGGTGTATCATCTGGAGAATGAGTGGCGCGTCTGAAGAAGCTGGTGGGCGAACGCTGAGGATCCCG **T6 p4**: 5′-CGGGATCCTCAGCGTTCGCCCACCAGCTTCTTCAGACGCGCCACTCATTCTCCAGATGATACACCTTGGACAGCAGTTCTTCCACC
T9	**T9 p1**: 5′-GGAATTCCATATGAAGCAGCTGGAAGATAAGGTGGAAGAGCTGCTGTCAAAGTGTACCATCTGGAAAACGAAGTGGCGCGTCTGAAGAAG **T9 p2**: 5′-CAGCAGTTCTTCCACCTTATCTTCCAGCTGCTTCATGTTCGCAATGTCATCTTTGGTCGCCATCAGCTTCTTCAGACGCGCCACTTC **T9 p3**: 5′-GATAAGGTGGAAGAACTGCTGTCCAAAGTGTACCATCTGGAAAACGAAGTGGCGCGTCTGAAGAAACTGGTGGGCGAACGCTGAGGATCCCG **T9 p4**: 5′-CGGGATCCTCAGCGTTCGCCCACCAGTTTCTTCAGACGCGCCACTTCGTTTTCCAGATGGTACACTTTGGACAGCAGTTCTTCCACCTTATC
A6	**A6 p1**: 5′-GGAATTCCATATGAAGCAACTTGAAGACAAAGTCGAAGAGCTTCTCTCAAGTTTATCATCTTGAGAACGAAGTTGCTCGTCTTAAG **A6 p2**: 5′-CCTTAGAAAGAAGTTCTTCGACCTTATCCTCAAGTTGCTTCATATCGCTTTGTCTCAATGAGTTTCTTAAGACGAGCAACTTCG **A6 p3**: 5′-CGAAGAACTTCTTTCTAAGGTTTACCATCTCGAAAATGAGGTTGTCGTTCAGAAGCTTGTTGGCGAACGCTGAGGATCCCG **A6 p4**: 5′-CGGGATCCTCAGCGTTCGCCAACAAGCTTCTTGAGACGAGCAACCCATTTCGAGATGGTAAACCTTAGAAAGAAGTTCTTCG
A7	**MP A6+K se**: 5′-CTTAAGAAACTCATTGAGAACAAGAAAGCCGATATGAAGCAAC **MP A6+K as**: 5′-GTTGCTTCATATCGGCTTTCTTGTTCTCAATGAGTTTCTTAAG
A9	**MP A6+KAD se**: 5′-CATTGAGAACAAAGCCGATAAGGCTGACATGAAGCAACTTGAGG **MP A6+KAD as**: 5′-CCTCAAGTTGCTTCATGTCAGCCTTATCGGCTTTGTTCTCAATG

DOI: 10.7554/eLife.11861.013

and A9 were constructed by site-directed mutagenesis using DNA fragment A6 as a template following the instructions of the QuikChange II XL Site-Directed Mutagenesis Kit. The DNA fragment coding for variant A9b was produced by gene synthesis (GenScript) and cloned in the NdeI and BamHI sites of pETHis1a_Nde1.

Protein expression and purification

A6, A7, A9, A9b, T6 and T9 were expressed in *E. coli* strain C41 (DE3), OMP100 and Tcar0761 constructs in XL1-blue. Cells were grown at 37°C until OD600 = 0.6, then expression was induced by addition of 1 mM isopropyl β-D-1-thiogalactopyranoside. Cells were cultivated for another 5 hr, harvested by centrifugation and disrupted using a French press cell (SLM Aminco). All proteins were purified under denaturing conditions. 6 M guanidinium chloride was added to the cell lysate and the sample stirred for 1 hr at room temperature. After centrifugation, the supernatant was loaded on a NiNTA column equilibrated with 20 mM Tris, pH 7.9, 400 mM NaCl, 10% glycerol, 6 M guanidinium chloride and bound proteins were eluted with a linear gradient of 0–0.5 M imidazol. Proteins were refolded by dialysis. Corresponding refolding buffers are listed in ***Table 1***. Refolded OMP100 was additionally subjected to a Superdex 75 column. For A6, A7, A9, A9b, T6 and T9 the N-terminal histidine tags were removed before crystallization. As the TEV cleavage site turned out to be not accessible for the TEV protease, the N-terminal tag was digested with Proteinase K. Subsequent analysis of the proteins by mass spectroscopy showed intact proteins lacking only the N-terminal extension including the histidine tag and the TEV cleavage site.

X-ray crystallography and structure analysis

Crystallization trials were set up in 96-well sitting-drop plates with drops consisting of 400 nl protein solution + 400 nl reservoir solution (RS) and reservoirs containing 75 µl RS. Crystallization and cryoprotection conditions for all crystal structures are listed in ***Table 4***. All crystals were loop mounted, flash frozen in liquid nitrogen, and all data collected at the SLS (Paul Scherrer Institute, Villigen,

Table 4. Crystallization and cryo condition.

Structure	Protein solution & concentration	Reservoir solution (RS)	Cryo solution
OMP 100	20 mM Tris pH 7.5, 150 mM NaCl, 3% (v/v) Glycerol, 3 mg/ml protein	0.1 M tri-Sodium citrate pH 5.5, 2% (v/v) Dioxane 15% (w/v) PEG 10,000	RS + 15% (v/v) PEG 400
A6	20 mM HEPES pH 7.2, 50 mM NaCl, 2% (v/v) Glycerol, 1 M Urea, 15 mg/ml protein	95 mM tri-Sodium citrate pH 5.6, 19% (v/v) Isopropanol, 19% (w/v) PEG 4000, 5% (v/v) Glycerol	-
A7	20 mM HEPES pH 7.3, 50 mM NaCl, 1 M Urea, 15 mg/ml protein	0.1 M Citric acid pH 3.5, 3 M NaCl	-
A9	20 mM HEPES pH 7.2, 50 mM NaCl, 2% (v/v) Glycerol, 1,5 M Urea, 17 mg/ml protein	1.6 M tri-Sodium citrate pH 6.5	-
A9b black	50 mM HEPES, 50 mM NaCl, 1 M Urea, 7.5 mg/ml protein	2.4 M Sodium malonate pH 5.0	-
A9b grey	50 mM HEPES, 50 mM NaCl, 1 M Urea, 7.5 mg/ml protein	0.2 M Sodium citrate, 0.1 M Bis Tris propane pH 6.5, 20% (w/v) PEG 3350	-
T6	20 mM HEPES pH 7.2, 50 mM NaCl, 1 M Urea, 13 mg/ml protein	0.2 M $CaCl_2$, 0.1 M HEPES pH 7.5, 30% (w/v) PEG 4000	-
$T9_6$	20 mM HEPES pH 7.2, 50 mM NaCl, 2% (v/v) Glycerol, 1.5 M Urea, 15 mg/ml protein	0.2 M Ammonium phosphate, 0.1 M TRIS pH 8.5, 50% (v/v) MPD	-
$T9_9$	„	0.1 M Citric acid pH 5.0, 20% (v/v) Isopropanol	RS + 1 M Urea +25% Glycerol
Tcar 0761	20 mM MOPS pH 7.2, 400 mM NaCl, 5% (v/v) Glycerol, 1.5 M Urea, 7 mg/ml protein	0.1 M tri-Sodium citrate pH 4.0, 30% (v/v) MPD	-

DOI: 10.7554/eLife.11861.014

Table 5. Data collection and refinement statistics.

Structure	OMP100	A6	A7	A9	A9b black	A9b grey	T6	$T9_6$	$T9_9$	Tcar0761
Beamline/Detector*	PXII / M	PXII / M	PXII / M	PXIII / M	PXII / P	PXII / P	PXII / P	PXII / M	PXIII / M	PXII / P
Wavelength (Å)	0.9786	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Trimers/AU	1	1	1/3	1	1	2	1	1	1	1/3
Space group	C2**	C2**	P321	$P2_1$	$P2_1$	$P2_1$	$P2_1$	$P2_1$	C2**	$P6_3$
a (Å)	62.1	60.4	38.2	65.2	26.2	71.1	34.2	25.1	60.8	37.9
b (Å)	35.9	34.8	38.2	34.6	37.5	35.0	27.0	38.3	35.1	37.9
c (Å)	198.5	104.2	87.1	67.5	95.0	106.2	101.0	105.0	112.2	179.2
β (°)	96.0	101.1	90	117.7	92.6	101.7	93.9	93.3	100.4	90
Resolution range (Å)***	32.9–2.30 (2.44–2.30)	30.0–2.10 (2.23–2.10)	18.2–1.37 (1.45–1.37)	33.7–1.80 (1.91–1.80)	34.9–1.35 (1.43–1.35)	38.1–2.00 (2.12–2.00)	34.1–1.60 (1.70–1.60)	34.9–1.80 (1.91–1.80)	19.5–2.00 (2.12–2.00)	32.3–1.60 (1.69–1.60)
Completeness (%)	92.4 (86.5)	97.3 (96.2)	99.0 (98.6)	98.9 (97.4)	95.9 (92.1)	92.4 (98.9)	98.2 (96.1)	97.1 (95.4)	98.7 (97.5)	99.2 (96.9)
Redundancy	2.84 (2.52)	3.71 (3.71)	6.35 (6.33)	3.70 (3.67)	3.72 (3.47)	3.29 (3.31)	3.04 (2.89)	3.94 (3.81)	3.73 (3.73)	3.69 (3.65)
I/σ(I)	14.0 (1.88)	15.5 (2.28)	18.2 (2.52)	14.3 (2.07)	17.6 (2.10)	13.9 (2.43)	13.6 (2.33)	14.5 (2.14)	19.5 (2.25)	20.3 (2.23)
R_{merge} (%)	4.2 (44.8)	4.8 (62.1)	5.1 (75.5)	5.1 (61.7)	3.4 (66.6)	5.0 (51.5)	4.4 (42.3)	7.2 (71.7)	4.0 (63.2)	2.9 (60.2)
R_{cryst} (%)	22.5	20.8	19.5	20.6	16.3	20.6	17.4	18.7	21.1	17.7
R_{free} (%)	25.4	25.1	23.8	25.6	19.9	25.3	20.5	22.6	25.5	21.3
PDB code	5APP	5APQ	5APS	5APT	5APU	5APV	5APW	5APX	5APY	5APZ

*M = MARRESEARCH mar225 CCD detector; P = DECTRIS PILATUS 6M detector

**twinned with apparent H32 symmetry and twinning operators
1/2*h-3/2*k,-1/2*h-1/2*k,-1/2*h+1/2*k-l and 1/2*h+3/2*k,1/2*h-1/2*k,-1/2*h-1/2*k-l

***values in parenthesis refer to the highest resolution shell

DOI: 10.7554/eLife.11861.015

Switzerland) under cryo conditions at 100 K using the beamlines and detectors indicated in **Table 5**. Data were processed and scaled using XDS (**Kabsch, 1993**). Structures were solved by molecular replacement using MOLREP (**Vagin and Teplyakov, 2000**). For OMP100 and the GCN4-fusion constructs, trimmed models of the SadAK3 structure (2WPQ) were used as search models. For Tcar0761, fragments of the T6 and A9 structures were used. After rebuilding with ARP/WARP (**Perrakis et al., 1999**), all structures were completed in cyclic manual modeling with Coot (**Emsley and Cowtan, 2004**) and refinement with REFMAC5 (**Murshudov et al., 1999**). Analysis with Procheck (**Laskowski et al., 1993**) showed excellent geometries for all structures. Data collection and refinement statistics are summarized in **Table 5**. Periodicity plots were calculated based on the output of TWISTER (**Strelkov and Burkhard, 2002**). Molecular depictions were prepared using MolScript (**Kraulis, 1991**), Raster3D (**Merritt and Bacon, 1997**) and Pymol (Schrödinger, LLC, New York, NY).

Bioinformatics

Sequence similarity searches were carried out at the National Institute for Biotechnology Information (NCBI; http://blast.ncbi.nlm.nih.gov/) and in the MPI Bioinformatics Toolkit (http://toolkit.tuebingen.mpg.de; **Biegert et al., 2006**), using PSI-Blast (**Altschul et al., 1997**) at NCBI and PatternSearch, CS-Blast (**Biegert and Söding, 2009**), HMMER (**Eddy, 2011**), HHblits (**Remmert et al., 2011**) and HHpred (**Söding et al., 2005**) in the MPI Toolkit. The sequence relationships of proteins identified in these searches were explored by clustering them according their pairwise Blast P-values in CLANS (**Frickey and Lupas, 2004**). Sequence logos were created from representative, non-redundant alignments using the WebLogo3 web server (**Crooks et al., 2004**) with composition correction switched off.

Secondary structure propensity was evaluated in the MPI Toolkit with the meta-tools Quick2D and Ali2D, and coiled-coil propensity was estimated with COILS/PCOILS (**Lupas et al., 1991**; **Gruber et al., 2006**) and MARCOIL (**Delorenzi and Speed, 2002**).

Searches for structures containing β-layers were performed over the Protein Data Bank (PDB, Dec 8 2015) in a two-step procedure: First, their torsion angles were scanned with seven-residue sliding windows of βββαααα and αααααβββ, where α must satisfy -70° $\leq \psi \leq$ -10° and -180° $\leq \varphi \leq$ -40°, and β must satisfy 20° $\leq \psi \leq$ 180° and -180° $\leq \varphi \leq$ -40°. Second, the central β residue of putative β-layer strands was required to form backbone hydrogen bonds (N-O distance $\leq$ 3.5 Å) to the equivalent residue of another β-layer strand within a biological assembly. All matches were verified by visual inspection. These searches were complemented by extensive interactive analyses of fibrous proteins in PDB.

Acknowledgements

We thank Reinhard Albrecht, Kerstin Baer and Silvia Deiss for technical assistance and are very grateful to the staff of beamline X10SA/Swiss Light Source for excellent technical support. This work was supported by the German Science Foundation (SFB 766, TP B4) and by institutional funds of the Max Planck Society.

Additional information

Funding

Funder	Grant reference number	Author
Max-Planck-Gesellschaft		Marcus D Hartmann Claudia T Mendler Jens Bassler Ioanna Karamichali Oswin Ridderbusch Andrei N Lupas Birte Hernandez Alvarez
Deutsche Forschungsgemeinschaft	SFB766 (TP B4)	Andrei N Lupas Birte Hernandez Alvarez

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

MDH, BHA, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; CTM, JB, IK, OR, Acquisition of data, Analysis and interpretation of data; ANL, Conception and design, Analysis and interpretation of data, Drafting or revising the article

Author ORCIDs

Marcus D Hartmann, http://orcid.org/0000-0001-6937-5677
Andrei N Lupas, http://orcid.org/0000-0002-1959-4836

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389–3402. doi: 10.1093/nar/25.17.3389

Alvarez BH, Gruber M, Ursinus A, Dunin-Horkawicz S, Lupas AN, Zeth K. 2010. A transition from strong right-handed to canonical left-handed supercoiling in a conserved coiled-coil segment of trimeric autotransporter adhesins. *Journal of Structural Biology* **170**:236–245. doi: 10.1016/j.jsb.2010.02.009

Bassler J, Hernandez Alvarez B, Hartmann MD, Lupas AN. 2015. A domain dictionary of trimeric autotransporter adhesins. *International Journal of Medical Microbiology* **305**:265–275. doi: 10.1016/j.ijmm.2014.12.010

Biegert A, Mayer C, Remmert M, Soding J, Lupas AN. 2006. The MPI bioinformatics toolkit for protein sequence analysis. *Nucleic Acids Research* **34**:W335–W339. doi: 10.1093/nar/gkl217

Biegert A, Soding J. 2009. Sequence context-specific profiles for homology searching. *Proceedings of the National Academy of Sciences of the United States of America* **106**:3770–3775. doi: 10.1073/pnas.0810767106

Bogomolovas J, Simon B, Sattler M, Stier G. 2009. Screening of fusion partners for high yield expression and purification of bioactive viscotoxins. *Protein Expression and Purification* **64**:16–23. doi: 10.1016/j.pep.2008.10.003

Brown JH, Cohen C, Parry DA. 1996. Heptad breaks in alpha-helical coiled coils: stutters and stammers. *Proteins* **26**:134–145. doi: 10.1002/(SICI)1097-0134(199610)26:2<134::AID-PROT3>3.0.CO;2-G

Chen J, Skehel JJ, Wiley DC. 1999. N- and c-terminal residues combine in the fusion-pH influenza hemagglutinin HA2 subunit to form an n cap that terminates the triple-stranded coiled coil. *Proceedings of the National Academy of Sciences* **96**:8967–8972. doi: 10.1073/pnas.96.16.8967

Crick F. 1953a. The fourier transform of a coiled-coil. *Acta Crystallographica* **6**:685–689.

Crick F. 1953b. The packing of alpha-helices: simple coiled-coils. *Acta Crystallographica* **6**:689–697.

Crooks G. 2004. WebLogo: a sequence logo generator. *Genome Research* **14**:1188–1190.

Deiss S, Hernandez Alvarez B, Bär K, Ewers CP, Coles M, Albrecht R, Hartmann MD. 2014. Your personalized protein structure: andrei n. lupas fused to GCN4 adaptors. *Journal of Structural Biology* **186**:380–385. doi: 10.1016/j.jsb.2014.01.013

Delorenzi M, Speed T. 2002. An Hmm model for coiled-coil domains and a comparison with Pssm-based predictions. *Bioinformatics* **18**:617–625. doi: 10.1093/bioinformatics/18.4.617

Eddy SR, Pearson W. 2011. Accelerated profile HMM searches. *PLoS Computational Biology* **7**:e1002195. doi: 10.1371/journal.pcbi.1002195

Emsley P, Cowtan K. 2004. *Coot* : model-building tools for molecular graphics. *Acta Crystallographica Section D Biological Crystallography* **60**:2126–2132. doi: 10.1107/S0907444904019158

Frickey T, Lupas A. 2004. Clans: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**:3702–3704. doi: 10.1093/bioinformatics/bth444

Grin I, Hartmann MD, Sauer G, Hernandez Alvarez B, Schütz M, Wagner S, Madlung J, Macek B, Felipe-Lopez A, Hensel M, Lupas A, Linke D. 2014. A trimeric lipoprotein assists in trimeric autotransporter biogenesis in enterobacteria. *Journal of Biological Chemistry* **289**:7388–7398. doi: 10.1074/jbc.M113.513275

Gruber M. 2003. Historical review: another 50th anniversary – new periodicities in coiled coils. *Trends in Biochemical Sciences* **28**:679–685. doi: 10.1016/j.tibs.2003.10.008

Gruber M, Söding J, Lupas AN. 2006. Comparative analysis of coiled-coil prediction methods. *Journal of Structural Biology* **155**:140–145. doi: 10.1016/j.jsb.2006.03.009

Hartmann MD, Ridderbusch O, Zeth K, Albrecht R, Testa O, Woolfson DN, Sauer G, Dunin-Horkawicz S, Lupas AN, Alvarez BH. 2009. A coiled-coil motif that sequesters ions to the hydrophobic core. *Proceedings of the National Academy of Sciences of the United States of America* **106**:16950–16955. doi: 10.1073/pnas.0907256106

Hartmann MD, Grin I, Dunin-Horkawicz S, Deiss S, Linke D, Lupas AN, Hernandez Alvarez B. 2012. Complete fiber structures of complex trimeric autotransporter adhesins conserved in enterobacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**:20907–20912. doi: 10.1073/pnas.1211872110

Hartmann MD, Dunin-Horkawicz S, Hulko M, Martin J, Coles M, Lupas AN. 2014. A soluble mutant of the transmembrane receptor Af1503 features strong changes in coiled-coil periodicity. *Journal of Structural Biology* **186**:357–366. doi: 10.1016/j.jsb.2014.02.008

Hernandez Alvarez B, Hartmann MD, Albrecht R, Lupas AN, Zeth K, Linke D. 2008. A new expression system for protein crystallization using trimeric coiled-coil adaptors. *Protein Engineering Design and Selection* **21**:11–18. doi: 10.1093/protein/gzm071

Hicks MR, Walshaw J, Woolfson DN. 2002. Investigating the tolerance of coiled-coil peptides to nonheptad sequence inserts. *Journal of Structural Biology* **137**:73–81. doi: 10.1006/jsbi.2002.4462

Hoiczyk E, Roggenkamp A, Reichenbecher M, Lupas A, Heesemann J. 2000. Structure and sequence analysis of yersinia YadA and moraxella UspAs reveal a novel class of adhesins. *The EMBO Journal* **19**:5989–5999. doi: 10.1093/emboj/19.22.5989

Huang P-S, Oberdorfer G, Xu C, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, Baker D. 2014. High thermodynamic stability of parametrically designed helical bundles. *Science* **346**:481–485. doi: 10.1126/science.1257481

Joh NH, Wang T, Bhate MP, Acharya R, Wu Y, Grabe M, Hong M, Grigoryan G, DeGrado WF. 2014. De novo design of a transmembrane Zn2+-transporting four-helix bundle. *Science* **346**:1520–1524. doi: 10.1126/science.1261172

Kabsch W. 1993. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Journal of Applied Crystallography* **26**:795–800. doi: 10.1107/S0021889893005588

Kraulis P. 1991. Molscript: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* **24**:946–950.

Laskowski R, MacArthur M, Moss D, Thornton J. 1993. Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**:283–291.

Leo JC, Lyskowski A, Hattula K, Hartmann MD, Schwarz H, Butcher SJ, Linke D, Lupas AN, Goldman A. 2011. The structure of e. coli IgG-binding protein d suggests a general model for bending and binding in trimeric autotransporter adhesins. *Structure* **19**:1021–1030. doi: 10.1016/j.str.2011.03.021

Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. *Science* **252**:1162–1164. doi: 10.1126/science.252.5009.1162

Lupas A. 1996. Coiled coils: new structures and new functions. *Trends in Biochemical Sciences* **21**:375–382. doi: 10.1016/S0968-0004(96)10052-9

Lupas AN, Gruber M. 2005. The structure of alpha-helical coiled coils. *Advances in Protein Chemistry* **70**:37–38. doi: 10.1016/S0065-3233(05)70003-6

Merritt EA, Bacon DJ. 1997. Raster3D: photorealistic molecular graphics. *Methods in Enzymology* **277**:505–524.

Murshudov G, Vagin A, Lebedev A, Wilson K, Dodson E. 1999. Efficient anisotropic refinement of macromolecular structures using fft. *Acta Crystallographica Section D Biological Crystallography* **55**:247–255.

Parry DA. 2014. Fifty years of fibrous protein research: a personal retrospective. *Journal of Structural Biology* **186**:320–334. doi: 10.1016/j.jsb.2013.10.010

Perrakis A, Morris R, Lamzin VS. 1999. Automated protein model building combined with iterative structure refinement. *Nature Structural Biology* **6**:458–463. doi: 10.1038/8263

Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative protein sequence searching by Hmm-Hmm alignment. *Nature Methods* **9**:173–175. doi: 10.1038/nmeth.1818

Strelkov SV, Burkhard P. 2002. Analysis of alpha-helical coiled coils with the program TWISTER reveals a structural mechanism for stutter compensation. *Journal of Structural Biology* **137**:54–64. doi: 10.1006/jsbi.2002.4454

Szczesny P, Lupas A. 2008. Domain annotation of trimeric autotransporter adhesins–daTAA. *Bioinformatics* **24**:1251–1256. doi: 10.1093/bioinformatics/btn118

Soding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**:W244–W248. doi: 10.1093/nar/gki408

Thomson AR, Wood CW, Burton AJ, Bartlett GJ, Sessions RB, Brady RL, Woolfson DN. 2014. Computational design of water-soluble α-helical barrels. *Science* **346**:485–488. doi: 10.1126/science.1257452

Vagin A, Teplyakov A. 2000. An approach to multi-copy search in molecular replacement. *Acta Crystallographica Section D Biological Crystallography* **56**:1622–1624. doi: 10.1107/S0907444900013780

Woolfson Dn. 2005. The design of coiled-coil structures and assemblies. *Advances in Protein Chemistry* **70**:79–112. doi: 10.1016/S0065-3233(05)70004-8

MCUR1 is the prototype of a protein family conserved in prokaryotes and eukaryotic organelles

Jyoti Adlakha, Ioanna Karamichali, Juthaporn Sangwallek[1], Silvia Deiss, Kerstin Bär, Murray Coles, Marcus D. Hartmann, Andrei N. Lupas, Birte Hernandez Alvarez*

Department of Protein Evolution, Max Planck Institute for Developmental Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany

[1]Present address:

Faculty of Medicine and Public Health, HRH Princess Chulabhorn College of Medical Science, Chulabhorn Royal Academy, 54 Kamphaeng Phet 6, Talat Bang Khen, Lak Si, Bangkok 10210, Thailand

* Correspondence:

B. Hernandez Alvarez,

Tel. –49 7071 601 356,

Fax –49 7071 601 349,

birte.hernandez@tuebingen.mpg.de

Summary

Membrane-bound coiled-coil proteins are important mediators of signaling, fusion, and scaffolding. Here, we delineate a heterogeneous group of trimeric membrane-anchored proteins in prokaryotes and eukaryotic organelles. They exhibit a characteristic head-neck-stalk-anchor architecture, in which a membrane-anchored coiled-coil stalk projects different N-terminal head domains via a β-layer neck. Based on sequence analysis, we identify different types of head domains and determine crystal structures of two representatives, the archaeal protein Kcr-0859 and the human CCDC90B, which possesses the most widespread head type. Using mitochondrial calcium uniporter regulator 1 (MCUR1), the functionally characterized paralog of CCDC90B, we study the role of individual domains, and find that the head interacts directly with the mitochondrial calcium uniporter (MCU) and is destabilized upon Ca^{2+} binding. Our data provide structural details of a class of membrane-bound coiled-coil proteins and identify the conserved head domain of the most widespread type as a mediator of their function.

Key words: crystal structure, coiled coil, β-layer, calcium, membrane, protein, mitochondria

Introduction

Coiled coils are ubiquitous structural units of proteins (Lupas et al., 2017). They may occur alone or as domains within larger proteins and fulfil many different functions (Hartmann, 2017; Lupas, 1996; Truebestein and Leonard, 2016). Coiled coils are, for instance, essential components of the cellular skeleton and of motor proteins that contribute to cellular motility (Squire et al., 2017). They act as molecular spacers, bridge defined molecular distances and function as levers in vesicle tethering, membrane fusion and chromosome segregation (Matityahu and Onn, 2018; Witkos and Lowe, 2017). Furthermore, since coiled coils are highly versatile in their interactions and oligomerization states, they facilitate ion transport and signal transduction across membranes, as components of channels and receptors (Gushchin and Gordeliy, 2018; Lupas and Bassler, 2017; Oxenoid et al., 2016).

Canonical coiled coils are highly regular, left-handed bundles of two or more α-helices that are wound around each other. Their amino acid sequence shows a characteristic heptad repeat pattern (*abcdefg*) with hydrophobic residues in positions *a* and *d*, sidechains of which pack in a regular geometry known as knobs-into-holes to form the core of the bundle. With an arrangement of seven residues over two helical turns, canonical coiled coils have a periodicity of 3.5 residues per turn. Naturally occurring coiled coils frequently show insertions of one, three, or four residues in their repeats. These types of insertions can be easily tolerated within coiled-coil structure, but cause local changes in periodicity and core packing (Brown et al., 1996; Lupas and Gruber, 2005). Insertions of three residues reduce the number of residues per turn, resulting in overwinding of the left-handed supercoil (Hartmann et al., 2009), and insertions of four residues increase the periodicity, leading to straightening of the coiled coil or even reversal of its handedness. Insertions of two or six residues overstrain the single helices and cause them to break. In trimeric coiled coils, this local overstrain is compensated for by the formation of β-layers, which are triangular supersecondary structural elements composed of three residues (Hartmann et al., 2016). β-Layers arise at the site of insertion and result in the continuation of the individual chains as short β-strands. The central β-layer residues form a characteristic inter-chain hydrogen bond network, which can coordinate a water molecule at the center of the bundles. β-Layers are either embedded within coiled coils or located at their ends, where they facilitate the

transition to structurally different domains. They occur as single elements but also repetitively arranged, forming α/β-coiled coils.

β-Layers are widely distributed in protein families, from phages and bacteria up to eukaryotes (Hartmann et al., 2016). One of these families, listed as DUF1640 in Pfam (Finn et al., 2016), comprises membrane-bound coiled-coil proteins conserved in prokaryotes and eukaryotic organelles. The domain organization of DUF1640 homologs follows a head-neck-stalk-anchor architecture with the head-stalk junction invariantly mediated by a β-layer neck. Human mitochondrial calcium uniporter regulator 1 (MCUR1) and its yeast homolog FMP32 represent the only family members that have been functionally characterized so far. Both proteins act as scaffold factors in the assembly of large inner membrane complexes in mitochondria, such as the Ca^{2+} uniporter channel complex and cytochrome c oxidase (COX) (Chaudhuri et al., 2016; Mallilankaraman et al., 2012; Paupe et al., 2015; Tomar et al., 2016).

Here, we present the results of a comprehensive search for integral membrane proteins characterized by a head-neck-stalk-anchor architecture, in which the neck is formed by a β-layer. We analyze the head domains of the identified proteins with regard to phylogeny and sequence conservation, and describe a new family of membrane-attached coiled-coil proteins widespread in prokaryotes and universal in mitochondria. We substantiate our bioinformatic data with crystal structures of the archaeal protein Kcr-0859 from *Candidatus Korarchaeum cryptofilum* and human CCDC90B. Based on these structures, we present homology models of human MCUR1 and of prokaryotic family members. In a functional study we analyze the role of single domains using MCUR1 as a model.

Results

The mempromCC family

In the course of investigating coiled-coil segments with irregularities within their repeat patterns, we identified the β-layer as a supersecondary structural element (Hartmann et al., 2016). Searches for β-layers in structurally characterized and uncharacterized proteins uncovered a characteristic sequence motif ([aliphatic]–A–T–

K–[polar]–[DE]), which we found in many unrelated protein families, including DUF1640. Within these we identified a class of proteins, again including DUF1640, which was characterized by a tripartite head-stalk-anchor architecture, with the head domains separated from the stalks by a β-layer-containing neck (Fig. 1A).

Aiming at a comprehensive bioinformatic characterization of this heterogeneous class of proteins, we used the head domains and succeeding β-layer neck sequences to identify more proteins by means of motif-based searches. The identified sequences were analyzed with regard to similarity, secondary structure, and domain architecture using different alignment and prediction tools, yielding a final set of 1085 integral membrane proteins from 270 species of all three domains of life and five viruses. All of these proteins are chromosomally encoded and constitute membrane-anchored homotrimers, as determined by the β-layer neck. They are characterized by a common head-neck-stalk-anchor architecture, but are highly diverse in sequence and differ remarkably in size. The majority of identified proteins is between 60 and 320 residues in length, with the longest representatives found in the fungal divisions *Ascomycota* and *Basidiomycota*, and the shortest examples in the bacterial genus *Xylella*. The head domains comprise of between 40 and 220 residues. Similarly, the coiled-coil stalks also vary in length and repeat patterns, even between proteins of closely related species. We find left-handed stalks that are entirely composed of canonical heptad repeats, but also segments built from repeats of eleven or fifteen residues, indicative of right-handed coiled-coils.

In order to analyze the phylogeny of the obtained sequences, we clustered their head-neck regions in CLANS, based on the strength of their all-against-all pairwise sequence similarities (see Methods). The resulting map shows many tightly connected central subclusters, containing about 70% of the input sequences, as well as several smaller subclusters at the periphery (Fig. 1B). The peripheral clusters exclusively contain prokaryotic sequences belonging to the same or closely related genera. The majority of the peripheral clusters are not connected at a P-value cutoff of 1.0 e-3, with only the groups of *Enterobacterales* and *Pseudomonas* as well as those of *Sulfurihydrogenibium* and the archaeal *Methanocaldococcus* showing sequence relationships to each other (Fig. 2).

The central subclusters comprise proteobacterial proteins of the α-, β-, γ- and δ-division and eukaryotic homologs localized in the organelles of metazoa, higher

plants, fungi and some algae. With the exception of some more broadly distributed fungal proteins, most of the sequences cluster closely together. The alignment of the head-neck segments originating from the central subclusters shows the conservation of a prominent FDT motif at the beginning of the head domain and of several other, more widely spaced residues (Fig. 2). Based on secondary structure prediction, the conserved head domain comprises two α-helical segments, which are separated by up to five residues and followed by a β-layer neck of the conserved sequence motif [aliphatic]–A–T–K–[polar]–[DE]. The bacterial sequences of the central subclusters show connections to the more distantly related groups of *Pseudomonas*, *Delftia* and *Chromatiales*. Although the head domains of these groups lack the FDT motif, they clearly share sequence similarity and predicted secondary structure with the central subclusters (Fig. 2). Based on these findings, we classify the proteins of the central subclusters and the most closely related peripheral subclusters of *Pseudomonas*, *Delftia* and *Chromatiales*, as belonging to a new protein family, mempromCC (membrane-attached proteins of prokaryotes and mitochondria containing coiled coils). Members of this family are expected to form homotrimers that display a head-neck-stalk-anchor architecture, with a C-terminal membrane-anchored coiled-coil stalk projecting the N-terminal head. The evolutionarily conserved head domain merges into the β-layer neck segment, which invariably mediates the head-stalk transition. The stalks of mempromCC homologs are highly variable in sequence and length. While prokaryotes contain one mempromCC homolog that completely lacks any N-terminal extensions or signal sequences, the majority of eukaryotic organisms has two paralogs that contain an N-terminal mitochondrial targeting signal (MTS), followed by intrinsically disordered segments of variable length.

In a next step, we set out to substantiate our bioinformatic analysis with structural data. Various proteins belonging to different subgroups of the cluster map and fragments thereof were overexpressed, purified and set up for crystallization. While most of the crystallization trials remained unsuccessful, we could obtain diffracting crystals for the cytosolic part of Kcr-0859 (WP_012309502) from *Candidatus Korarchaeum cryptofilum* OPF8, an archaeal protein of unknown function, and a large fragment covering the head, neck, and beginning of the stalk from the human mempromCC homolog CCDC90B (Fig. 3).

Crystal structure of crenarchaeal Kcr-0859

Kcr-0859ΔTM, a construct of the archaeal protein Kcr-0859 lacking the C-terminal membrane anchor (Fig. 3A), was expressed in *E. coli* C41, purified to homogeneity and set up for crystallization. Analysis of Kcr-0859ΔTM by Far-UV CD spectroscopy shows a typical α-helical spectrum with characteristic minima at 208 nm and 221 nm. The protein unfolds with a melting temperature of Tm ≈ 50 °C (Fig. S1A). The best diffracting crystals yielded a dataset to a resolution of 2.2 Å and could be solved via SAD-phasing. In agreement with our expectation, the Kcr-0859ΔTM structure shows an elongated parallel trimer comprising of a helical head domain, which is connected via a β-layer neck to a coiled-coil stalk (Fig. 3B). Each monomer of the head bundle consists of three short helices (α1, α2, α3), which are connected by short turns and arranged perpendicular to each other. Within the trimer, helices are packed in parallel with helix α2 of each monomer accommodated between α1` and α3`` of the other chains. The central axis of the head domain is kinked by 40 degrees relative to the coiled-coil axis of the adjacent stalk. This kinked arrangement is likely a result of crystal packing constraints, but is also indicative of a certain degree of flexibility of the neck region, as reported for β-layer-mediated transitions in DALL domains of trimeric autotransporter adhesins (TAAs) (Hartmann et al., 2012; Koiwai et al., 2016). Formed by the first three residues of the motif MATKED, the β-layer directly succeeds the α3 helices and forms the characteristic inter-chain hydrogen bonds between the central alanine residues.

Following the neck, the structure shows 76 residues of the 96-residue cytosolic part of the coiled-coil stalk. This whole region does not possess noticeable supercoiling over large extents, as anticipated from its annotation (Fig. S1B), which indicated an overall sequence periodicity of 3.64 residues per turn, only marginally different from the 3.63 residues expected for an undistorted α-helix. Towards its C-terminal end, the stalk contains two YxD motifs, both of which are resolved in the structure. These polar motifs are commonly found in right-handed coiled coils, where they convey structural specificity and stability by forming inter-chain hydrogen bonds between the hydroxyl groups of the tyrosines and the carboxyl groups of the aspartates (Alvarez et al., 2010).

Crystal structure of human CCDC90B

For the structural characterization of the mempromCC family, we initially focused on the human paralogs MCUR1 and CCDC90B. While both proteins interact with components of the mitochondrial Ca^{2+} uniporter channel complex, only MCUR1 essentially regulates mitochondrial Ca^{2+} uptake by modulating uniporter activity (Chaudhuri et al., 2016; Mallilankaraman et al., 2012; Sancak et al., 2013; Tomar et al., 2016).

For both proteins (Fig. S1B, C), we designed different constructs covering the head and neck segments, and varying in the length and nature of the stalk included. We succeeded to solubly express $CCDC90B_{43\text{-}125}$, a fragment comprising the head-neck segment and the first 21 residues of the coiled-coil stalk, in *E. coli* ArcticExpress (DE3) (Fig. 3A). The coiled-coil fragment of the construct was fused in-register to a $GCN4_{N16V}$ adaptor, which forms a mixture of dimers and trimers in solution (Harbury et al., 1993). In the past, we have frequently used such in-register fusions of GCN4 variants to stabilize coiled-coil fragments for structural characterization (Deiss et al., 2014; Hernandez Alvarez et al., 2008).

Far-UV CD spectroscopy shows that $CCDC90B_{43\text{-}125}$-$GCN4_{N16V}$ adopts an α-helical structure with characteristic spectral minima at 208 nm and 220 nm (Fig. S2A). The fusion protein melts with a midpoint of Tm = 71°C. As thermal denaturation of the $GCN4_{N16V}$ adaptor is expected to take place at Tm > 95°C (Knappenberger et al., 2002), it seems likely that the melting curve shows the thermal unfolding of the $CCDC90B_{43\text{-}125}$ fragment and the $GCN4_{N16V}$ segment unfolds in a second step at higher temperatures. The trimeric oligomerization state of the protein was verified by SEC-MALS, where it eluted as a single peak with an apparent molecular weight of 41.5 kDa, corresponding three times the theoretical molecular mass of 13.5 kDa (Fig. S2B).

In crystallization screens, we obtained crystals for which we collected a dataset to a resolution of 2.1 Å and which we could solve via molecular replacement using trimeric coiled-coil segments as search models. The crystal structure of $CCDC90B_{43\text{-}125}$-$GCN4_{N16V}$ shows an extended helical trimer (Fig. 3C). The solved structure starts at D62, the central residue of the conserved FDT motif at the beginning of the head domain. The first 19 residues, corresponding to positions 43 - 61 of CCDC90B, are not resolved in the electron density. This can be attributed to the flexibility of the N-terminal region in the crystal, as the integrity of the protein was

confirmed by analyzing crystals on SDS-PAGE. The head domain is formed by helices α1 (residues 62-72) and α2 (residues 77-101). Overall, it is distinct from the head of the crenarchaeal protein *Kcr-0859* and can be described as an antiparallel six-helix bundle extending into a trimeric coiled-coil segment, which connects directly to the β-layer neck. In the six-helix bundle, the α1-helices are packed antiparallel against the α2-helices of the same and a neighboring chain, stabilized by a hydrophobic interface. At the tip of the head, the residues connecting α1 to α2 form a cap-like structure, stabilizing the trimer with hydrophobic contacts of the conserved F75 and a network of intra- and interchain hydrogen bonds between the side chain of residue Q79 and the backbone amide and carboxyl moieties of neighboring residues D76 and G74, respectively (Fig. 3D).

The C-terminal part of the α2-helices of each chain (residues 87-101), which is not in contact with the α1-helices, is composed of 15 residues and accommodated as a pentadecad over four helical turns in a right-handed coiled coil. The β-layer neck comprises the first three residues of the motif MVTQAQ (residues 102-107) with the central β-layer residue V103 forming the inter-chain backbone hydrogen bonds and coordinating a water molecule in the center (Fig. 3E). The β-layer is followed by a segment of the natural stalk comprising of a hendecad and a heptad, which shows a transition from a slightly right-handed to a left-handed coiled coil and merges continuously into the $GCN4_{N16V}$ adaptor.

Homology model of MCUR1

The human paralogs CCDC90B and MCUR1 share a sequence identity of 57% in the region covering the head, neck, stalk and membrane anchor. The individual domains of the paralogs are identical in length, with the stalks even sharing the same repeat pattern. The N-terminal extensions preceding the head domains, however, differ remarkably in length and sequence, representing the most distinctive feature of the paralogs (Fig. 3A). Based on secondary structure prediction, they comprise of about 60 residues in CCDC90B and 160 residues in MCUR1, and include a mitochondrial signal peptide, followed by an intrinsically disordered segment (Oates et al., 2013; Zimmermann et al., 2017).

For structural characterization of MCUR1, we designed multiple expression constructs comprising the conserved head domain, but lacking the N-terminal disordered extension. A fusion construct of $MCUR1_{160-230}$ with $GCN4_{N16V}$, equivalent to the successfully utilized $CCDC90B_{43-125}$-$GCN4_{N16V}$ construct, was found to be well expressed in the insoluble fraction but could not be refolded. The same fragment lacking a C-terminal fusion, $MCUR1_{160-230}$ (Fig. 3A), could be refolded successfully in physiological buffer, but did not yield any diffracting crystals.

Next, we assessed the suitability of $MCUR1_{160-230}$ for solution NMR spectroscopy. The 1D ^{1}H spectrum showed adequate dispersion and diffusion coefficients consistent with an oligomer, most likely the expected trimer. However, experiments using isotope-labelled samples to acquire spectra for sequential assignment indicated that only a stretch of residues, corresponding to the core of the head domain (C173-S190), showed sharp signals with good dispersion (Fig S3). Other residues were much broader with low dispersion, could not be definitively assigned or were not observed. Under these circumstances, it was not possible to proceed to a high-resolution structure.

Facing these difficulties, we constructed a homology model of MCUR1 using the $CCDC90B_{43-125}$-$GCN4_{N16V}$ crystal structure as template. Lacking the flexible N-terminal extension preceding the head, the MCUR1 model mirrors CCDC90B in length as well as in domain architecture (Fig. 3F). However, the charge distribution of the head domains differs significantly for the two paralogs (Fig. 4). The electrostatic surface of $CCDC90B_{43-125}$ displays a prominent acidic patch at the top of the head, with residues D69, E71, D76 and E81 contributing the bulk of negative charge. While the corresponding glutamate residues E176 and E186 are conserved in MCUR1, the aspartates are replaced by non-polar residues L174 and A181. Residues E176 and D177 of MCUR1 primarily contribute to the small negatively charged patches on top of the MCUR1 head.

The MCUR1 head domain interacts with the N-terminal domain of mitochondrial calcium uniporter

In order to characterize the function of the single domains of mempromCC homologs, we focused on MCUR1, which represents the functionally best characterized member of the family. Several studies identified MCUR1 as an interactor of the N-terminal

domain of the mitochondrial calcium uniporter (MCU-NTD) (Lee et al., 2015; Tomar et al., 2016). In order to identify the regions of MCUR1 involved in this interaction, we designed different constructs in which single domains were either substituted or deleted in order to identify the regions of MCUR1 mediating binding to MCU (Fig. 5A). Following co-expression of C-terminal FLAG-tagged MCUR1 variants with MCU containing a C-terminal HA-tag in HEK293T cells, samples were subjected to immunoprecipitation with an anti-FLAG antibody and subsequently analyzed by western blotting (Fig. 5B). Expectedly, MCUR1 was able to pull down MCU. Similarly, MCU binds to MCUR1-Δβ, a construct lacking the β-layer neck motif MVTKMQ, indicating the dispensability of the β-layer neck for the interaction. MCUR1-GCN4pII, in which the natural stalk was replaced by a fragment comprising thirteen heptads derived from the sequence of the trimeric GCN4pII variant (Harbury et al., 1993), is also not impaired in MCU binding. However, deletion of almost the entire stalk in MCUR1-Δstalk reduces MCU binding significantly and concurrently decreases protein stability, suggesting the length of the coiled-coil stalk to be critical for the interaction. MCUR1-ΔDR, lacking the disordered region, was similarly expressed at a low level, but co-precipitated MCU. In contrast, MCUR1-ΔDR-Head, lacking the disordered region including the conserved head domain, does not bind to MCU at all, showing that the head domain is essential for the MCUR1-MCU interaction. Based on the observation that paralogous CCDC90B also interacts with MCU (Tomar et al., 2016), we designed MCUR1-Head90B to verify this interaction. This MCUR1 variant, in which the native head was exchanged by the head domain of CCDC90B, binds MCU, albeit with slightly decreased affinity.

MCUR1 is processed by proteases; this is seen from the immunoprecipitation experiments and was also observed by others before (Chaudhuri et al., 2016; Tomar et al., 2016). To narrow down the region undergoing proteolytic cleavage, we expressed MCUR1-FLAG in HEK293 cells. Performing immunoprecipitation with an anti-FLAG antibody, the samples were separated by SDS-PAGE and the three most prominent protein bands were cut from the gel, digested with trypsin and analysed by mass spectrometry (Fig. S4). The two larger bands, with an apparent molecular weight of 41 kDa and 37 kDa, were identified as full-length MCUR1 with uncleaved and cleaved MTS, respectively. The smaller, most prominent band, with an apparent molecular weight of 25 kDa, was recognized as a fragment of MCUR1 that lacks the

first 140 residues. This includes almost the entire disordered region and suggests that MCUR1 function is proteolytically controlled.

Next, we analyzed the interaction of the head domains of MCUR1 and CCDC90B with the MCU-NTD *in vitro* by Microscale thermophoresis (MST). MCU-NTD possesses a β-grasp-like fold and represents the soluble, matrix-localized part of the channel protein. It is connected via a coiled-coil domain (CC1) with the membrane-spanning part and regulates channel assembly and activity in dependence of Ca^{2+} and Mg^{2+} (Lee et al., 2016; Lee et al., 2015; Oxenoid et al., 2016; Yoo et al., 2018). For *in vitro* binding studies, we expressed MCU_{75-233} comprising MCU-NTD including CC1. The purified protein was well folded, showing CD spectra of mixed α-helical and β-stranded secondary structure (Fig. S5). Analyzing MCU_{75-233} by SDS-PAGE, we find the protein fragment to form a protein ladder with single bands corresponding to different oligomeric states, similar as reported previously for full-length MCU (Lee et al., 2016). On Blue Native PAGE, we observe that Ca^{2+} concentrations ≥ 0.5 mM shift the equilibrium towards higher order oligomers. This was rather surprising as Ca^{2+} was found to destabilize MCU-NTD and to promote its disassembly (Lee et al., 2016). As the construct used in this study additionally includes CC1, this implicates this domain to mediate Ca^{2+}-dependent oligomerization. The interaction of MCU_{75-233} with the head domains of MCUR1 and CCDC90B were analyzed by MST and dissociation constants of K_d = 12.7 ± 3.5 µM for $MCUR1_{160-230}$ and K_d = 58.7 ± 1.6 µM for $CCDC90B_{43-125}$ were determined (Fig. 5C). Concentrations of up to 1 mM $CaCl_2$ did not show any significant effect on the interaction (Fig. 5D). Measurements at higher Ca^{2+} concentrations were unfeasible due to higher order oligomer formation and aggregation observed for both, MCU_{75-233} and $MCUR1_{160-230}$.

The MCUR1 head domain is destabilized by Ca^{2+}

In the following we examined, whether MCUR1, as a regulator of mitochondrial calcium uptake, is itself affected by Ca^{2+}. Assessing oligomer formation and stability by SEC-MALS, we found $MCUR1_{160-230}$ to elute as a single peak with a calculated molecular mass of 22.6 kDa (Fig. 6A). Given the theoretical mass of 8.3 kDa for a monomer, this corresponds to a trimer in solution. Addition of Ca^{2+} did not have any visible effect on oligomerization. MST was further used to analyze potential Ca^{2+} binding by titration of the fluorescently labelled head domains of MCUR1 and

CCDC90B with $CaCl_2$ in a concentration range of 1 µM – 40 mM. Whereas $CCDC90B_{43-125}$ does not show any affinity for Ca^{2+}, the curve monitored for $MCUR1_{160-230}$ clearly showed Ca^{2+} binding (Fig. 6B). We determined affinity constants for divalent cation binding to the MCUR1 head with $K_d \approx 0.6$ mM for $CaCl_2$ and $K_d \approx 5.0$ mM for $MgCl_2$ (Fig. 6C). At concentrations > 5 mM $CaCl_2$, $MCUR1_{160-230}$ started to precipitate. This effect was observed to be less pronounced in the presence of Mg^{2+}.

Next, we analyzed possible effects of Ca^{2+} on secondary structure by CD spectroscopy. In the absence of Ca^{2+}, $MCUR1_{160-230}$ shows a typical α-helical spectrum with characteristic minima at 208 nm and 220 nm and unfolds upon heating with a melting temperature of Tm = 71 °C (Fig. 6D, E). Monitoring single CD spectra at increasing temperatures for $CCDC90B_{43-125}$ and $MCUR1_{160-230}$ in Ca^{2+}-free buffer, we observed that thermal unfolding of $MCUR1_{160-230}$ α-helical structure is accompanied by the irreversible formation of soluble β-like structure, becoming visible as an emerging signal at 216 nm (Fig. 6F). In contrast, $CCDC90B_{43-125}$ is not affected (Fig. S6).

At higher Ca^{2+} concentrations, we did not find any effect on α-helical structure of $CCDC90B_{43-125}$ (Fig. S6) and $MCUR1_{160-230}$ (Fig. 6D). However, $MCUR1_{160-230}$ was strongly impaired in thermal stability with increasing Ca^{2+} concentrations, visible as a gradual reduction of the melting temperature (Fig. 6E). The lowered intensity of the CD signal at 216 nm results from increasing amounts of precipitate formed with rising Ca^{2+} concentrations during heat denaturation (Fig. 6G). This process could be inhibited upon addition of the Ca^{2+} chelator EGTA. In the presence of Mg^{2+}, the thermal stability of $MCUR1_{160-230}$ is equally lowered.

In addition to Ca^{2+}-induced precipitation during thermal unfolding, we observed formation of translucent gel-like precipitates while incubating $MCUR1_{160-230}$ at room temperature over time and during concentration of the protein. To analyze the nature of these higher order formations, samples of $MCUR1_{160-230}$ were incubated at different Ca^{2+} concentrations for several hours at 25 °C. As observed by visual inspection, cloudy aggregates formed in the presence of Ca^{2+} within 3 hours, whereas no precipitates were visible in the Ca^{2+}-free samples even after 24 hours. Analysis of the samples by Transmission Electron Microscopy (TEM) showed β-fibril formation of $MCUR1_{160-230}$ in a Ca^{2+}-dependent manner (Fig. 6H). In Ca^{2+}-free

samples, we observe only single short protofibrils. These form extended curvilinear protofibrillar structures with increasing Ca^{2+} concentrations up to mature amyloid fibrils in buffer containing 5 mM Ca^{2+}. The effect of Mg^{2+} was found to be less pronounced. Analyses of aggregation and amyloid propensity using computational prediction tools, such as Amylpred2, TANGO, and ZipperDB (Fernandez-Escamilla et al., 2004; Tsolis et al., 2013), identified a common amyloidogenic hotspot in the α2-helix of the MCUR1 head domain, including residues 184-196, which might promote fibril formation (Fig. S7). For the equivalent fragment of the CCDC90B head, which differs slightly in sequence, the score for β-fibrillar structure formation is comparatively low.

Model structure and localization of prokaryotic mempromCC homologs

Prokaryotic mempromCC homologs share the head-neck-stalk-anchor domain architecture with their eukaryotic counterparts and show high sequence similarity in the head domain. Based on this, we used the $CCDC90B_{43-125}$ crystal structure to construct models of two homologous *Caulobacter* proteins, namely YP_002517927 from *C. crescentus* NA1000 and WP_018113394 from *C. sp.* JGI 0001013-D04 (Fig. 7A). Showing the conserved overall domain arrangement, the models clearly display the natural variability of prokaryotic stalk domains with respect to their length and the number of β-layers. The stalk of the *C. crescentus* NA1000 protein is considerably shorter than the stalk of its *Caulobacter sp.* JGI 0001013-D04 homolog, which contains eight successive β-layers.

Prokaryotic mempromCC homologs do not contain any signal peptide for export across the membrane and are therefore likely to be anchored to the inner bacterial membrane via their C-terminal transmembrane helices. To verify this experimentally, we analyzed the cellular localization of the *Caulobacter* mempromCC homolog CCNA_02554 (YP_002517927), referred in the following as MpcC (mempromCC of *Caulobacter*), using TEM and cell fractionation. MpcC was fused to an N-terminal HA-tag and expressed at low levels in *C. crescentus* NA1000 from the high-copy plasmid pBXMCS4 utilizing the leakiness of the xylose-inducible promoter (Thanbichler et al., 2007). Cryosections of *Caulobacter* cells expressing HA-MpcC were stained with polyclonal anti-HA antibody, followed by immunogold labelled secondary antibody. TEM micrographs show the majority of the electron-dense

particles to accumulate in close proximity to the inner side of the cytoplasmic membrane (Fig. 7B), supporting the assumption that the protein is C-terminally anchored to the bacterial cytoplasmic membrane pointing the head into the cytoplasm.

Cellular localization was further studied by subcellular fractionation of *Caulobacter* cells expressing MpcC. Analysis of the samples by western blotting using polyclonal anti-MpcC antiserum clearly shows that MpcC is enriched in the inner membrane fraction (Fig 7C). The purity of the subcellular fractions was judged by probing the blot with antibodies specific for the inner membrane protein TimA and the outer membrane protein CpaC.

In order to examine the membrane topology of MpcC, we performed Proteinase K treatment of spheroblasts prepared from *Caulobacter* cells expressing MpcC fused to an N-terminal HA-tag and three C-terminal FLAG-epitopes [HA-MpcC-$(FLAG)_3$]. Fractionated spheroblasts were analyzed by western blotting using anti-HA and anti-FLAG antibodies (Fig. 7D). In control samples, both antibodies equally recognize two bands in the inner membrane fraction, a larger one corresponding to full length HA-MpcC-$FLAG_3$ and a smaller one that is apparently degraded from the C-terminus by cellular proteases. In Proteinase K-treated samples, the N-terminal HA-epitope is present in three bands: the biggest corresponding to full-length protein and the other two representing proteolytic fragments, about 2-4 kDa smaller in size. Probing this blot with anti-FLAG antibody, only the two larger bands were recognized. This suggests that the FLAG-tagged C-terminus of MpcC is accessible for digestion by Proteinase K and therefore located on the periplasmic side of the inner membrane. The incomplete digestion observed from multiple proteolytic fragments, can be attributed to the close proximity of the $(FLAG)_3$-tag to the membrane surface hampering the access for the protease. Our results thus demonstrate the localization of bacterial mempromCC homologs to the inner plasma membrane with an N in-C out orientation.

Discussion

In this paper, we present a comprehensive study of a heterogeneous group of trimeric integral membrane proteins from prokaryotes and eukaryotes, which share a

common head-neck-stalk-anchor architecture. The N-terminal helical head domain of these proteins is invariably connected to a C-terminal membrane-anchored coiled-coil stalk by one or multiple β-layer necks. Analyzing the evolutionary relationship of their head domains based on sequence similarity, we identify multiple subgroups of different head types. We substantiate our bioinformatic analysis with structural data obtained from crystal structures of the crenarchaeal protein Kcr-0859 and human CCDC90B, which belong to different head subgroups. The head domain of CCDC90B belongs to the biggest subgroup comprising more than two thirds of the head sequences identified in this study, including all from eukaryotes and many from proteobacteria. We have named this group the mempromCC family and clearly demarcate it from other proteins of similar domain architecture.

The stalk domains of prokaryotic mempromCC homologs are more diverse in sequence and length than their eukaryotic counterparts. The homology models of the human paralog MCUR1 and two *Caulobacter* proteins, constructed in this work, give an impression of their diversity in terms of length and the number of β-layer necks. We further show that prokaryotic mempromCC homologs, which lack any signal peptide, are anchored to the bacterial inner membrane via their C-terminal transmembrane helices, with the head projected by the stalk into the cytoplasm. Eukaryotic mempromCC homologs contain an organelle localization signal, which is in most cases predicted to target mitochondria. In previous studies, localization to the inner mitochondrial membrane has been verified experimentally for yeast FMP32 and the human paralogs MCUR1 and CCDC90B (Chaudhuri et al., 2016; Mallilankaraman et al., 2012; Paupe et al., 2015). However, mitochondrial localization does not seem to be universal, as the homolog Rat1 from *Chlamydomonas reinhardtii* was found to be associated with chloroplast membranes (Balczun et al., 2005). The majority of eukaryotic species contains two or more paralogs, which primarily differ in the N-terminal intrinsically disordered segments preceding their conserved head domains. This diversity among paralogous proteins may point to different cellular functions. At present, only few eukaryotic mempromCC homologs including two human paralogs have been functionally characterized. Chloroplast-located Rat1 was identified in a cDNA library screen for factors able to complement a mutant defective in *tsc*A splicing (Balczun et al., 2005). Deletion of *fmp*32 in yeast and knock down of MCUR1 in human fibroblasts produce comparable defects in COX assembly (Paupe et al., 2015). Paralogous MCUR1 and CCDC90B interact with each

other, but also with essential components of the mitochondrial Ca^{2+} uniporter channel complex, including the selective Ca^{2+} channel subunit MCU and the regulatory single-pass membrane protein EMRE. Although interacting with the same proteins, the paralogs differ in their biological activities: only MCUR1, but not CCDC90B, was shown to be essential for active MCU complex formation and cellular Ca^{2+} uptake (Chaudhuri et al., 2016; Mallilankaraman et al., 2012; Tomar et al., 2016).

Based on these studies, we used MCUR1 as a model to examine the functional significance of the individual domains of mempromCC homologs. With the knowledge of the exact domain boundaries, we identify the head domains of MCUR1 and CCDC90B to mediate direct binding of both proteins to the N-terminal domain of MCU. The stalk domain is not directly involved in MCU binding, but its length is critical for the interaction, by serving as a projector of the head. The role of the disordered region remains unclear from our experiments, as a variant lacking this segment is strongly impaired in stability. In agreement with previous studies (Chaudhuri et al., 2016; Tomar et al., 2016), we find that MCUR1 is processed *in vivo* and show that a major part of the disordered region of MCUR1 is cleaved. Both, full-length as well as the processed form of MCUR1, interact with MCU, but it remains to be shown, whether they are also functionally equivalent. It is possible that MCUR1 activity and stability are proteolytically regulated. Similarly, the level of non-assembled EMRE is strictly controlled by the mAAA proteases AFG3L2 and SPG7 during MCU complex formation (Tsai et al., 2017). Also other mitochondrial proteins, such as dynamin-like GTPase OPA1, a key regulator of mitochondrial dynamic, and Parkinson disease-related Ser/Thr kinase PINK1, are known to be under tight proteolytic control (Ali and McStay, 2018; Greene et al., 2012; MacVicar and Langer, 2016).

The paralogous head domains of CCDC90B and MCUR1 show high sequence identity, but clearly differ in biophysical properties like surface charge distribution and susceptibility to divalent cations. Whereas Ca^{2+} and Mg^{2+} do not show any measurable effect on the CCDC90B head domain, both ions bind to the MCUR1 head at concentrations in the single-digit millimolar range and significantly impair its thermal stability. Similar effects have been reported for the N-terminal domain of MCU, which is destabilized and affected in oligomerization upon Ca^{2+} binding (Lee et al., 2015). These differences in Ca^{2+} susceptibility correlate with the observation that both paralogous head domains interact with MCU, but only MCUR1 affects MCU-

mediated mitochondrial Ca^{2+} uptake *in vivo* (Tomar et al., 2016). In contrast to CCDC90B, unfolding of the MCUR1 head domain is accompanied by β-fibril formation and this process is accelerated in the presence of Ca^{2+} and Mg^{2+}. Similar modulating effects of Ca^{2+} on β-fibril formation are described for the neurological disease-related proteins α-synuclein and the amyloid beta peptide (Aβ) (Han et al., 2018; Isaacs et al., 2006), both of which accumulate as amyloid structures at mitochondrial membranes causing organelle dysfunction (Kawamata and Manfredi, 2017). As present experiments were performed with a single domain separated from its natural domain context, it remains unclear whether Ca^{2+} exerts the same effects on MCUR1 secondary structure *in vivo*. However, considering the spatial proximity of MCUR1 to the calcium-releasing side of the uniporter channel, a functional relevance for Ca^{2+} binding in the regulation of MCUR1 seems conceivable. Future studies will be required to map potential Ca^{2+} binding sites of the MCUR1 head domain and to assess their significance for MCUR1 stability and activity *in vivo*.

In summary, we present here a combined bioinformatic and structural study of a heterogeneous group of membrane-bound coiled coil proteins. These proteins share the same domain architecture, but, based on their sequence diversity, they are likely to participate in different cellular processes. Our data suggest that they fulfil their functions generally via their head domains, which are projected by the membrane-anchored stalk.

Acknowledgement

We thank Reinhard Albrecht for setting up crystallization experiments and are grateful to the macromolecular crystallography group at the SLS for excellent technical support. We thank Katharina Hipp and Matthias Floetenmeyer for their help in sample preparation and performance of EM experiments. We are grateful to Edgardo Sepulveda for his support in *Caulobacter* culturing and genetics, Mateusz Korycinski and Jens Bassler for their help in sequence analysis and modelling and Vikram Alva for critical reading of the manuscript. We thank Prof. Martin Thanbichler for providing pBXMCS4 and Prof. Lucy Shapiro for anti-CpaC antiserum. This work was supported by institutional funds from the Max Planck Society and by the German science foundation Grant SFB766/B4 to AL and BHA.

Author contribution

Conceptualization: A.L., B.H.A.

Methodology, Investigation: J.A., I.K., J.S., S.D., K.B., M.C., M.D.H., A.L., B.H.A.

Writing – Original Draft: J.A., B.H.A. with contributions from the other authors

Funding Acquisition: A.L., B.H.A.

Declaration of interests

The authors declare no competing interests.

References

Ali, S., and McStay, G.P. (2018). Regulation of Mitochondrial Dynamics by Proteolytic Processing and Protein Turnover. Antioxidants (Basel) *7*.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Alvarez, B.H., Gruber, M., Ursinus, A., Dunin-Horkawicz, S., Lupas, A.N., and Zeth, K. (2010). A transition from strong right-handed to canonical left-handed supercoiling in a conserved coiled-coil segment of trimeric autotransporter adhesins. J Struct Biol *170*, 236-245.

Anwari, K. (2012). Isolate and Sub-fractionate Cell Membranes from Caulobacter crescentus. Bio Protoc *2*.

Balczun, C., Bunse, A., Hahn, D., Bennoun, P., Nickelsen, J., and Kuck, U. (2005). Two adjacent nuclear genes are required for functional complementation of a chloroplast trans-splicing mutant from Chlamydomonas reinhardtii. Plant J *43*, 636-648.

Borchert, N., Dieterich, C., Krug, K., Schutz, W., Jung, S., Nordheim, A., Sommer, R.J., and Macek, B. (2010). Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. Genome Res *20*, 837-846.

Brown, J.H., Cohen, C., and Parry, D.A. (1996). Heptad breaks in alpha-helical coiled coils: stutters and stammers. Proteins *26*, 134-145.

Chaudhuri, D., Artiga, D.J., Abiria, S.A., and Clapham, D.E. (2016). Mitochondrial calcium uniporter regulator 1 (MCUR1) regulates the calcium threshold for the mitochondrial permeability transition. Proc Natl Acad Sci U S A *113*, E1872-1880.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr D Biol Crystallogr *62*, 1002-1011.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol *26*, 1367-1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res *10*, 1794-1805.

Deiss, S., Hernandez Alvarez, B., Bar, K., Ewers, C.P., Coles, M., Albrecht, R., and Hartmann, M.D. (2014). Your personalized protein structure: Andrei N. Lupas fused to GCN4 adaptors. J Struct Biol *186*, 380-385.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods *4*, 207-214.

Ely, B. (1991). Genetics of Caulobacter crescentus. Methods Enzymol *204*, 372-384.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr *60*, 2126-2132.

Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol *22*, 1302-1306.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A.*, et al.* (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res *44*, D279-285.

Franz-Wachtel, M., Eisler, S.A., Krug, K., Wahl, S., Carpy, A., Nordheim, A., Pfizenmaier, K., Hausser, A., and Macek, B. (2012). Global detection of protein kinase D-dependent phosphorylation events in nocodazole-treated human cells. Mol Cell Proteomics *11*, 160-170.

Frickey, T., and Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics *20*, 3702-3704.

Greene, A.W., Grenier, K., Aguileta, M.A., Muise, S., Farazifard, R., Haque, M.E., McBride, H.M., Park, D.S., and Fon, E.A. (2012). Mitochondrial processing peptidase regulates PINK1 processing, import and Parkin recruitment. EMBO Rep *13*, 378-385.

Gruber, M., Soding, J., and Lupas, A.N. (2006). Comparative analysis of coiled-coil prediction methods. J Struct Biol *155*, 140-145.

Gushchin, I., and Gordeliy, V. (2018). Transmembrane Signal Transduction in Two-Component Systems: Piston, Scissoring, or Helical Rotation? Bioessays *40*.

Han, J.Y., Choi, T.S., and Kim, H.I. (2018). Molecular Role of Ca(2+) and Hard Divalent Metal Cations on Accelerated Fibrillation and Interfibrillar Aggregation of alpha-Synuclein. Sci Rep *8*, 1895.

Harbury, P.B., Zhang, T., Kim, P.S., and Alber, T. (1993). A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science *262*, 1401-1407.

Hartmann, M.D. (2017). Functional and Structural Roles of Coiled Coils. Subcell Biochem *82*, 63-93.

Hartmann, M.D., Grin, I., Dunin-Horkawicz, S., Deiss, S., Linke, D., Lupas, A.N., and Hernandez Alvarez, B. (2012). Complete fiber structures of complex trimeric autotransporter adhesins conserved in enterobacteria. Proc Natl Acad Sci U S A *109*, 20907-20912.

Hartmann, M.D., Mendler, C.T., Bassler, J., Karamichali, I., Ridderbusch, O., Lupas, A.N., and Hernandez Alvarez, B. (2016). alpha/beta coiled coils. Elife *5*.

Hartmann, M.D., Ridderbusch, O., Zeth, K., Albrecht, R., Testa, O., Woolfson, D.N., Sauer, G., Dunin-Horkawicz, S., Lupas, A.N., and Alvarez, B.H. (2009). A coiled-coil motif that sequesters ions to the hydrophobic core. Proc Natl Acad Sci U S A *106*, 16950-16955.

Hernandez Alvarez, B., Hartmann, M.D., Albrecht, R., Lupas, A.N., Zeth, K., and Linke, D. (2008). A new expression system for protein crystallization using trimeric coiled-coil adaptors. Protein Eng Des Sel *21*, 11-18.

Isaacs, A.M., Senn, D.B., Yuan, M., Shine, J.P., and Yankner, B.A. (2006). Acceleration of amyloid beta-peptide aggregation by physiological concentrations of calcium. J Biol Chem *281*, 27916-27923.

Kabsch, W. (2010). Xds. Acta Crystallogr D Biol Crystallogr *66*, 125-132.

Kall, L., Krogh, A., and Sonnhammer, E.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res *35*, W429-432.

Kawamata, H., and Manfredi, G. (2017). Proteinopathies and OXPHOS dysfunction in neurodegenerative diseases. J Cell Biol *216*, 3917-3929.

Knappenberger, J.A., Smith, J.E., Thorpe, S.H., Zitzewitz, J.A., and Matthews, C.R. (2002). A buried polar residue in the hydrophobic interface of the coiled-coil peptide, GCN4-p1, plays a thermodynamic, not a kinetic role in folding. J Mol Biol *321*, 1-6.

Koiwai, K., Hartmann, M.D., Linke, D., Lupas, A.N., and Hori, K. (2016). Structural Basis for Toughness and Flexibility in the C-terminal Passenger Domain of an Acinetobacter Trimeric Autotransporter Adhesin. J Biol Chem *291*, 3705-3724.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567-580.

Lee, S.K., Shanmughapriya, S., Mok, M.C.Y., Dong, Z., Tomar, D., Carvalho, E., Rajan, S., Junop, M.S., Madesh, M., and Stathopulos, P.B. (2016). Structural Insights into Mitochondrial Calcium Uniporter Regulation by Divalent Cations. Cell Chem Biol *23*, 1157-1169.

Lee, Y., Min, C.K., Kim, T.G., Song, H.K., Lim, Y., Kim, D., Shin, K., Kang, M., Kang, J.Y., Youn, H.S*., et al.* (2015). Structure and function of the N-terminal domain of the human mitochondrial calcium uniporter. EMBO Rep *16*, 1318-1333.

Lupas, A. (1996). Coiled coils: new structures and new functions. Trends Biochem Sci *21*, 375-382.

Lupas, A.N., and Bassler, J. (2017). Coiled Coils - A Model System for the 21st Century. Trends Biochem Sci *42*, 130-140.

Lupas, A.N., Bassler, J., and Dunin-Horkawicz, S. (2017). The Structure and Topology of alpha-Helical Coiled Coils. Subcell Biochem *82*, 95-129.

Lupas, A.N., and Gruber, M. (2005). The structure of alpha-helical coiled coils. Adv Protein Chem *70*, 37-78.

MacVicar, T., and Langer, T. (2016). OPA1 processing in cell death and disease - the long and short of it. J Cell Sci *129*, 2297-2306.

Mallilankaraman, K., Cardenas, C., Doonan, P.J., Chandramoorthy, H.C., Irrinki, K.M., Golenar, T., Csordas, G., Madireddi, P., Yang, J., Muller, M.*, et al.* (2012). MCUR1 is an essential component of mitochondrial Ca2+ uptake that regulates cellular metabolism. Nat Cell Biol *14*, 1336-1343.

Matityahu, A., and Onn, I. (2018). A new twist in the coil: functions of the coiled-coil domain of structural maintenance of chromosome (SMC) proteins. Curr Genet *64*, 109-116.

Murshudov, G.N., Vagin, A.A., Lebedev, A., Wilson, K.S., and Dodson, E.J. (1999). Efficient anisotropic refinement of macromolecular structures using FFT. Acta Crystallogr D Biol Crystallogr *55*, 247-255.

Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztanyi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L.*, et al.* (2013). D(2)P(2): database of disordered protein predictions. Nucleic Acids Res *41*, D508-516.

Oxenoid, K., Dong, Y., Cao, C., Cui, T., Sancak, Y., Markhard, A.L., Grabarek, Z., Kong, L., Liu, Z., Ouyang, B.*, et al.* (2016). Architecture of the mitochondrial calcium uniporter. Nature *533*, 269-273.

Paupe, V., Prudent, J., Dassa, E.P., Rendon, O.Z., and Shoubridge, E.A. (2015). CCDC90A (MCUR1) is a cytochrome c oxidase assembly factor and not a regulator of the mitochondrial calcium uniporter. Cell Metab *21*, 109-116.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods *8*, 785-786.

Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods *9*, 173-175.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. Proteins *23*, 318-326.

Sancak, Y., Markhard, A.L., Kitami, T., Kovacs-Bogdan, E., Kamer, K.J., Udeshi, N.D., Carr, S.A., Chaudhuri, D., Clapham, D.E., Li, A.A.*, et al.* (2013). EMRE is an essential component of the mitochondrial calcium uniporter complex. Science *342*, 1379-1382.

Sheldrick, G.M. (2008). A short history of SHELX. Acta Crystallogr A *64*, 112-122.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res *33*, W244-248.

Squire, J.M., Paul, D.M., and Morris, E.P. (2017). Myosin and Actin Filaments in Muscle: Structures and Interactions. Subcell Biochem *82*, 319-371.

Thanbichler, M., Iniesta, A.A., and Shapiro, L. (2007). A comprehensive set of plasmids for vanillate- and xylose-inducible gene expression in Caulobacter crescentus. Nucleic Acids Res *35*, e137.

Thein, M., Sauer, G., Paramasivam, N., Grin, I., and Linke, D. (2010). Efficient subfractionation of gram-negative bacteria for proteomics studies. J Proteome Res *9*, 6135-6147.

Tokuyasu, K.T. (1973). A technique for ultracryotomy of cell suspensions and tissues. J Cell Biol *57*, 551-565.

Tomar, D., Dong, Z., Shanmughapriya, S., Koch, D.A., Thomas, T., Hoffman, N.E., Timbalia, S.A., Goldman, S.J., Breves, S.L., Corbally, D.P.*, et al.* (2016). MCUR1 Is a Scaffold Factor for the MCU Complex Function and Promotes Mitochondrial Bioenergetics. Cell Rep *15*, 1673-1685.

Truebestein, L., and Leonard, T.A. (2016). Coiled-coils: The long and short of it. Bioessays *38*, 903-916.

Tsai, C.W., Wu, Y., Pao, P.C., Phillips, C.B., Williams, C., Miller, C., Ranaghan, M., and Tsai, M.F. (2017). Proteolytic control of the mitochondrial calcium uniporter complex. Proc Natl Acad Sci U S A *114*, 4388-4393.

Tsolis, A.C., Papandreou, N.C., Iconomidou, V.A., and Hamodrakas, S.J. (2013). A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. PLoS One *8*, e54175.

Vagin, A., and Teplyakov, A. (2000). An approach to multi-copy search in molecular replacement. Acta Crystallogr D Biol Crystallogr *56*, 1622-1624.

Witkos, T.M., and Lowe, M. (2017). Recognition and tethering of transport vesicles at the Golgi apparatus. Curr Opin Cell Biol *47*, 16-23.

Yoo, J., Wu, M., Yin, Y., Herzik, M.A., Jr., Lander, G.C., and Lee, S.Y. (2018). Cryo-EM structure of a mitochondrial calcium uniporter. Science.

Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kubler, J., Lozajic, M., Gabler, F., Soding, J., Lupas, A.N., and Alva, V. (2017). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol.

Figure legends

Figure 1

Cluster map of the head-neck segments of membrane-bound β-layer containing proteins sharing a common head-neck-stalk-anchor architecture. (A) Schematic drawing displaying the domain architecture of proteins explored in the bioinformatic study comprising of a head domain, connected via a single or multiple β-layers with a coiled-coil stalk that passes in a C-terminal transmembrane helix (TM). **(B)** Cluster map of the head-neck segments of β-layer containing proteins identified by means of motif-based searches. The map was generated based on sequence similarities of the head-neck regions measured as BLAST P-values obtained by all-against-all pairwise sequence comparison. Clustering of sequences was performed with CLANS in 2D at a P-value cutoff of 1.0e-3 using default settings. Single sequences are depicted as circles colored according to their taxonomic classification. Evolutionary related sequences are connected by grey lines based on BLAST score values. Triangles and squares highlight proteins of particular interest and stars indicate proteins for which the crystal structure has been solved within the scope of this work. Homologs of the mempromCC family belong to the central subcluster and the peripheral subclusters PS and DL, which are connected by lines. Archaeal sequences form two clusters, one including sequences from *Methanocaldococcus* (MT) and one from Thaumarchaeota (T), Crenarchaeota (C) and Korarchaeota (K). Peripheral clusters are abbreviated as *Achromobacter* (AC), *Chromatiales* (CH), *Candidatus liberibacter* (CL), *Campylobacter* (CM), *Delfia* (DL), *Desulfovibrio* (DS), *Enterobacteriales* (EB), *Flavobacteriia* (FL), *Helicobacter* (HL), *Pseudomonas* (PS), *Sulfurihydrogenibium* (SL), *Sphingobacteria* (SP), *Thermoprotei* (TH), and *Xylella* (XL).

Figure 2

Sequence alignment of the head-neck segment of the mempromCC family and groups of *Pseudomonas*/*Enterobacteriales* (PS/EB) and *Crenarchaeota*. If multiple β-layers are present, only the first is included in the alignment. Multiple sequence alignments were generated using Clustal Omega. Secondary structure information was plotted on the alignment using Ali2D. Helical regions are shown in pink with color intensities correlating with increasing prediction rates. Conserved

residues are printed in bold. Blue letters indicate residues conserved in more than two thirds, black letters in at least half of the sequences. Hydrophobic positions, present in at least 50% of the sequences of each group, are marked as “h” and highlighted in gray. of Species are colored accordingly to their taxonomic classification (Fig. 1B) and abbreviated as follow: HS1 (NP_068597.2, CCDC90B, *Homo sapiens*), HS2 (NP_001026883.1, MCUR1, *Homo sapiens*), AT (NP_973473.1, *Arabidopsis thaliana*), CR (XP_001694431.1, Rat1, *Chlamydomonas reinhardtii*), RS (CCO26633.1, *Rhizoctonia solani AG-1 IB*), SC (Q05867.1, YL283, *Saccharomyces cerevisiae S288c*), SP (O14042.1 *Schizosaccharomyces pombe 972h-*), CC (YP_002517927.1, *Caulobacter crescentus NA1000*), CS (WP_018113394.1, *Caulobacter sp. JGI 0001013-D04*), CG (WP_006683118.1, *Candidatus Glomeribacter gigasporarum*), HP (WP_021111668.1, *Haemophilus parasuis*), PC (AIC21840.1, *Pseudomonas chlororaphis*), EH (WP_029762635.1, *Ectothiorhodospira haloalkaliphila*), CE (ETX04478.1, *Candidatus Entotheonella Sp. Tsy2*), PS (WP_016781459.1, *Pseudomonas fragi*), JA (CDG82950.1, *Janthinobacterium agaricidamnosum NBRC 102515 = DSM 9628*), PA (WP_014605518.1, *Pantoea ananatis*), EC (WP_001737198.1, *Escherichia coli*), SE (WP_000890813.1, *Salmonella enterica*), MY (ZP_09704076.1, *Metallosphaera yellowstonensis MK1*), SA (WP_024084599.1, *Sulfolobus acidocaldarius*), TA (WP_020962471.1, *Thermofilum adornatus*), PF (WP_014025983.1, *Pyrolobus fumarii*), KC (WP_012309502.1, Kcr-0859, *Candidatus Korarchaeum cryptofilum* OPF8).

Figure 3

Crystal structures of Kcr-0859 from *Candidatus Korarchaeum cryptofilum* OPF8 and human CCDC90B with the model of MCUR1. (A) Schematic representation of the domain architectures of Kcr-0859, CCDC90B and MCUR1 and the constructs Kcr-0859ΔTM, $CCDC90B_{43\text{-}125}$-$GCN4_{N16V}$ and $MCUR1_{160\text{-}230}$ used for structural studies. Single domains include mitochondrial targeting signal sequence (MTS), the disordered region (DR), the head domain (head), the stalk domain (stalk) and the transmembrane helix (TM). The β-layer neck is shown in red. **(B)** Cartoon representation of the trimeric crystal structure of Kcr-0859ΔTM. The β-layer is highlighted in gray. **(C)** Cartoon representation of the trimeric crystal structure of

CCDC90B$_{43-125}$-GCN4$_{N16V}$ with the GCN4$_{N16V}$ fusion shown in gray. **(D)** Detailed view of the top of the CCDC90B head domain showing conserved F75 and the interchain and intrachain interaction network formed by residues G74, D76 and Q79. **(E)** Close up of the β-layer neck of CCDC90B, highlighting the hydrogen bonds between the central β-layer residues V103 coordinating a water molecule. **(F)** Cartoon presentation of the homology model of MCUR1 spanning the head-neck-stalk region (residues 167-336). The N-terminal disordered regions of the three monomers are not included in the model and drawn as dotted lines. Proteins Kcr-0859 and MCUR1 are shown to be C-terminally anchored to the membrane with the head-neck-stalk region localized in the cytosol and matrix, respectively.

Figure 4

Surface charge of the head domains of MCUR1 and CCDC90B. The panels show side views (lower panels) of MCUR1 (residues 167-230) and CCDC90B (residues 62-125) and top views (upper panels) of their head domains. Negative charges are colored in red and positive charges in blue. Negatively charged residues are labeled.

Figure 5

The MCUR1 head domain interacts directly with the N-terminal domain of MCU. **(A)** Schematic representation of MCUR1 constructs used in pull down assays. **(B)** HEK293 cells were transfected with MCU-HA and different MCUR1-FLAG variants. Immunoprecipitation (IP) from cell lysates was performed using anti-FLAG antibody. Whole lysates and IP samples were analyzed by western blotting with anti-FLAG and anti-HA antibodies. **(C)** MST experiments to analyze binding of MCUR1 and CCDC90B to MCU-NTD. MCUR1$_{160-230}$ and CCDC90B$_{43-125}$ were titrated against fluorescently labelled MCU$_{75-233}$. **(D)** MST experiment measuring the effect of Ca^{2+} on the interaction of MCUR1$_{160-230}$ with MCU$_{75-233}$. Fluorescently labelled MCU$_{75-233}$ was titrated against MCUR1$_{160-230}$ in the absence or presence of 0.1 mM and 1 mM $CaCl_2$ concentrations.

MST experiments described in panels (C) and (D) were performed with Alexa Fluor 647 red-NHS dye labelled MCU$_{75-233}$ at a concentration of 20 nM. Single data points,

representing means of three measurements with standard error bars, were plotted and fitted for determination of dissociation constants (K_d).

Figure 6

The head domain of MCUR1 is destabilized upon divalent cation binding. (A) SEC MALS data of $MCUR1_{160-230}$ in the presence of 1 mM EGTA and 5 mM $CaCl_2$. Calculated molecular masses are indicated. **(B)** MST measurements analyzing divalent cation binding to the head domains of CCDC90B and MCUR1. **(C)** Titration curves to analyze binding of the CCDC90B and MCUR1 heads to $CaCl_2$ and $MgCl_2$ in a range of 1 µM – 40 mM. Single data points are means of three replicates with standard error and were plotted for determination of affinity constants (K_d). **(D)** Far-UV CD spectra of $MCUR1_{160-230}$ at 20 ˚C in the presence of EGTA, $MgCl_2$ and $CaCl_2$ at indicated concentrations. For the yellow curve (1 mM $CaCl_2$ + 1 mM EGTA), $CaCl_2$ was added to the sample and chelated by adding equal molar amounts of EGTA before measurement. **(E)** Thermal melting curves for $MCUR1_{160-230}$ in the presence of EGTA, $MgCl_2$ or $CaCl_2$ measured at 208 nm. Concentrations and color codes correspond to (C). Calculated melting temperatures (Tm) are indicated. **(F)** Thermal melting curves for $MCUR1_{160-230}$ measured at 216 nm. **(G)** Single Far-UV CD spectra of $MCUR1_{160-230}$ measured at different temperatures in the presence of EGTA, $CaCl_2$ and $MgCl_2$. **(H)** TEM micrographs showing β-fibril formation of $MCUR1_{160-230}$ in dependence of EGTA, $CaCl_2$ and $MgCl_2$ following incubation at 25˚C for 24 hours.

Figure 7

mempromCC homologs *from Caulobacter species*. (A) Cartoon presentation of structural models encompassing the entire head-neck-stalk region of the mempromCC homologs YP_002517927 from *C. crescentus* NA1000 and WP_018113394 from *C. sp.* JGI 0001013-D04. β-layer necks are shown in red, head domains in green and stalks in blue. **(B)** Electron micrograph showing localization of HA-MpcC expressed in *C. crescentus* NA1000 cells. Cryo-sections were stained with anti-HA antibody followed by immunogold labelled secondary antibody and analyzed by TEM. **(C)** Analysis of subcellular localization of HA-MpcC expressed in

C. crescentus NA1000 cells. Whole cell lysate (lysate), fractions of the outer (OM) and inner membrane (IM) and the soluble fraction (sol. fract., cytosol and periplasm) were analyzed by western blotting using anti-MpcC antibody. Purity of the fractions was assessed using antibodies against marker proteins TimA for IM and CpaC for OM fractions. **(D)** Analysis of membrane topology of HA-MpcC-$(FLAG)_3$ expressed in *C. crescentus* NA1000. Spheroblasts were incubated with Proteinase K and subjected to fractionation. Samples of whole cell lysate, OM, IM and cytosol were analyzed by western blotting using anti-HA and anti-FLAG antibodies.

Table 1. Data collection and refinement statistics

	Kcr-0859ΔTM	**CCDC90B$_{43-125}$-GCN4$_{N16V}$**
Wavelength (Å)	0.979	1.07
Space group	$P2_12_12$	$P2_12_12$
Cell dimensions (Å)	a=197.5 b=48.9, c=51.1	a=36.0 b=298.2 c=29.3
Monomers / ASU	3	3
Resolution range data collection (Å)	39.3 - 2.19 (2.32 – 2.19)	37.3 - 2.10 (2.23 – 2.10)
Completeness (%)	99.0 (95.7)	99.8 (98.8)
Redundancy	3.40 (3.27)	6.27 (6.46)
I/σ(I)	9.23 (1.74)	13.4 (1.89)
R_{merge} (%)	8.6 (58.5)	8.9 (86.7)
CC(1/2)	99.8 (79.8)	99.9 (83.5)
Resolution range refinement (Å)	39.3 - 2.19 (2.25 – 2.19)	37.3 - 2.10 (2.15 - 2.10)
R_{cryst} (%)	24.3 (31.7)	23.1 (38.6)
R_{free} (%)	27.3 (33.3)	25.7 (40.7)
RMSD Bond angles / lengths	1.04 / 0.0065	1.34 / 0.012
Ramachandran statistics (%)	100 / 0 / 0	97.8 / 2.2 / 0
PDB accession code	**6H9L**	**6H9M**

Values in parentheses refer to the highest resolution shell. The Ramachandran statistics show the percentage of residues in favored / allowed / other regions.

Methods

Bioinformatic analysis

Sequence similarity searches were started with head-neck sequences of proteins we had identified previously as having a head-neck-stalk-anchor architecture (Hartmann et al., 2016), using PSI-Blast (Altschul et al., 1997) at the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/), and HHblits (Remmert et al., 2011) and HHpred (Soding et al., 2005) in the MPI Bioinformatics Toolkit (Zimmermann et al., 2017). Newly identified sequences were added to the dataset and used as starting points for further searches.

Analyses of identified sequences were performed using multiple prediction tools, including Quick2D and Ali2D (Zimmermann et al., 2017), and Pcoils (Gruber et al., 2006) in the MPI Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de/#/), and Phobius (Kall et al., 2007), TMHMM (Krogh et al., 2001), and SignalP (Petersen et al., 2011) in their respective home servers.

For sequence clustering, head-neck segments were analyzed according to their pairwise Blast P-values (Altschul et al., 1997) using CLANS (Frickey and Lupas, 2004) at default parameter settings. Multiple sequence alignments were generated with Clustal Omega (Sievers et al., 2011).

Cloning

For recombinant expression of proteins in *E. coli*, the corresponding DNA fragments were codon optimized and synthesized by gene synthesis. Fragments coding for Kcr-0859ΔTM, corresponding to residues 1-136 of Kcr-0859 (WP_012309502), $MCU_{75\text{-}233}$ (NP_612366), $MCUR1_{160\text{-}230}$ (NP_001026883), $CCDC90B_{43\text{-}125}$ (NP_068597), $CCDC90B_{43\text{-}125}$-$GCN4_{N16V}$ were cloned in pETHis_1a (G. Stier, EMBL Heidelberg) for overexpression with an N-terminal 6xHis-tag cleavable by TEV protease.

DNA fragments encoding tagged variants of MpcC (CCNA_02554, YP_002517927), were synthesized by gene synthesis and cloned in pBXMCS4 (Thanbichler et al., 2007). For expression of HA-MpcC-$(FLAG)_3$ and HA-MpcC, the plasmids were transformed in *C. crescentus* NA1000 by electroporation (Ely, 1991).

For transient expression in HEK293 cells, DNA fragments encoding MCU-HA (full length MCU with C-terminal HA-tag), FLAG-MCUR1 (full length MCUR1 with N-

terminal FLAG-tag) and MCUR1 variants fused to an N-terminal FLAG-tag were synthesized by gene synthesis and cloned in vector pCDNA3.1. The MCUR1-Δβ construct lacks residues 207-212, encoding the β-layer neck motif MVTKMQ. In MCUR1-GCN4pII, the fragment 231-321 was replaced by the same number of residues of sequence $(MKQIEDKIEEILSKIYHIENEIARIKKL)_3$-MKQIEDK derived from the trimeric GCN4pII variant (Harbury et al., 1993). Residues 224-321, 43-159 and 43-233 were deleted in constructs MCUR1-Δstalk, MCUR1-ΔDR and MCUR1-ΔDR-Head, respectively. In MCUR1-Head90B, the fragment spanning residues 162-234 of MCUR1 was replaced by residues 45-129 of the head region of CCDC90B.

Protein expression and purification

Recombinant Kcr-0859ΔTM and MCU_{75-233} were overexpressed in *E. coli* C41 strain. $CCDC90B_{43-125}$-$GCN4_{N16V}$, $CCDC90B_{43-125}$ and $MCUR1_{160-230}$ were expressed in *E. coli* ArcticExpress *(DE3)* cells.

E. coli strains were grown in Luria broth (LB) supplemented with kanamycin at 37˚C for C41 and ArcticExpress (DE3). Protein expression was induced with 1 mM isopropyl β-D-thiogalactoside (IPTG) at an optical density of OD_{600} = 0.5. Following incubation for 4 h at 37 °C for *E. coli* C41 and 24 h at 12 °C for ArcticExpress (DE3), cells were harvested by centrifugation. The cell pellet was resuspended in lysis buffer containing 20 mM Tris, pH 7.5, 150 mM NaCl, 4 mM $MgCl_2$, DNaseI, 1 mM phenylmethylsulfonyl fluoride (PMSF) and cOmplete EDTA-free Protease Inhibitor Cocktail (Roche), and subsequently lysed by sonication.

Kcr-0859ΔTM, $CCDC90B_{43-125}$-$GCN4_{N16V}$ and $CCDC90B_{43-125}$ were purified under native conditions. Following centrifugation of the cell lysate to remove cell debris, the supernatant was loaded on a Ni-NTA Agarose column pre-equilibrated with buffer A (20 mM Tris, pH 7.6, 150 mM NaCl). Bound proteins were eluted with a two-step gradient including a step of 5% buffer B (20 mM Tris, pH 7.6, 500 mM NaCl, 0.5 M imidazole) followed by linear gradient of 5-100% buffer B. Protein containing fractions were dialyzed against buffer A and incubated with TEV protease for His-tag cleavage. Cleaved protein was separated from the 6xHis-tagged TEV protease and proteolytic fragments, by re-loading the sample on Ni-NTA. Fractions containing the cleaved protein were pooled and purified to homogeneity by gel filtration on Superdex 75.

MCU_{75-233} and $MCUR1_{160-230}$ were purified under denaturing conditions. Cell lysate was stirred in 6 M guanidine hydrochloride (Gua-HCl) at room temperature for 1 h. Following centrifugation, the supernatant was loaded on Ni-NTA Agarose column equilibrated with 20 mM Tris, pH 8.0, 300 mM NaCl, 6 M Gua-HCl, and bound proteins were eluted with a linear gradient of 0-0.5 M imidazole in the same buffer. Purified proteins were refolded by dialysis against buffer containing 20 mM Tris, pH 7.6, 150 mM NaCl and 1 mM DTT. Following cleavage by TEV protease, the protein was purified to homogeneity performing a second Ni-NTA column followed by gel filtration on Superdex 75, as described above.

For structural characterization of $MCUR1_{160-230}$ by NMR, *E. coli* C41 cells were grown in M9 minimal medium supplemented with ^{13}C-glucose and ^{15}N-ammonium chloride. Protein expression was induced at OD_{600} = 0.6 with 1 mM IPTG. Following incubation at 20 ˚C for 18 h, cells were harvested by centrifugation. Protein purification was performed under denaturing conditions as described above.

CD spectroscopy

Circular dichroism (CD) spectra were recorded using a Jasco J-810 spectropolarimeter equipped with a JASCO-423S Peltier Controller. CD measurements were performed at a protein concentration of 0.5 mg/ml for $MCUR1_{160-230}$, and 0.2 mg/ml for $CCDC90B_{43-125}$-$GCN4_{N16V}$ and $CCDC90B_{43-125}$ in 10 mM Tris, pH 7.5, 20 mM NaCl using a cuvette with a path length of 1 mm. Single CD spectra were recorded at a speed of 100 nm/min with a data pitch of 0.5 and a response time of 1 s. Each spectrum represents the average of five scans corrected by the signal of buffer scan. Thermal melting curves were recorded by monitoring ellipticity at indicated wavelengths in a temperature range from 10 – 95 ˚C applying a ramp of 0.5 ˚C/min. Blank correction, smoothing of data and calculation of molecular ellipticities and melting temperatures were performed using Spectra Manager Software (JASCO).

MST

In microscale thermophoresis (MST) binding experiments a dilution series of a ligand was titrated against fluorescently labelled protein. For labelling, 100 µl of 20 µM

protein sample in 20 mM HEPES, pH 7.6, 150 mM NaCl were mixed with 100 µl of 60 µM Alexa Fluor 647 red-NHS amine-reactive dye solution in DMSO and incubated for 1 h in the dark. For desalting, a NanoTemper Gravity Flow Column B was equilibrated with MST buffer (50 mM Tris, pH 7.6, 150 mM NaCl and 0.05% Tween-20) and the reaction mixture was loaded on the column followed by 300 µl MST buffer. Samples were eluted with 600 µl MST buffer and stored as 10 µl aliquots at -80 ˚C. Following optimization of the concentration of labelled protein, serial dilutions of non-fluorescent ligand were prepared in MST buffer and mixed with the labelled protein in a 1:1 ratio. Tubes were incubated for 15 min, centrifuged at 13,000 rpm for 10 min to remove aggregates and filled in Monolith NT 'Premium coated' capillaries. Measurements were performed using MicroScale Thermophoresis instrument Monolith NT.115 (NanoTemper Technologies) with in-built MO.Control software. Data were analyzed with MO.Affinity Analysis software.

SEC-MALS

Size Exclusion Chromatography and Multi Angle Light Scattering (SEC-MALS) experiments were performed to calculate the absolute molar mass of proteins and their oligomeric states in solution using a 1260 Infinity II HPLC (Agilent) coupled to a miniDawn TREOS and Optilab T-rEX refractive index detector (Wyatt Technologies). Proteins were applied at a concentration of 5 mg/ml on an AdvanceBio SEC 130 Å column (for MCUR1 constructs) or an AdvanceBio SEC 300 Å column (for MCUR1 constructs) equilibrated with 20 mM Tris, pH 7.5, 150 mM NaCl and 0.2% NaN_3 and separated at a flow-rate of 0.5 ml/min at 18 °C. For analyses of $MCUR1_{160-230}$, 1 mM TCEP was added to the buffer. Data analysis and molecular mass calculation was performed using ASTRA software package (Wyatt Technologies).

Mammalian cell culture and transfection

HEK293 cells were cultivated in DMEM containing 10% FBS and 2 mM glutamine. Cells were transiently transfected with pCDNA3.1 plasmid encoding MCU-HA and MCUR1-FLAG variants using Lipofectamine 2000 according to the manufacturer´s protocol (Thermo Fisher Scientific).

Immunoprecipitation

HEK293 cells were harvested 20-24 h post-transfection and washed twice with ice-cold PBS. The cell pellet was resuspended in lysis buffer containing 20 mM HEPES, pH 7.3, 150 mM NaCl, 10 mM KCl, 1% Nonidet P-40, 8% glycerol, 1 mM PMSF and cOmplete EDTA-free Protease Inhibitor Cocktail (Roche) and incubated for 10 min on ice. Following sonication, cell debris were pelleted in a microcentrifuge at 13,000 rpm. The soluble fraction was incubated with Anti-FLAG M2 Magnetic Beads for 2 h at 4 °C. The beads were collected and washed three times in lysis buffer using a magnetic separator. Proteins bound to the beads were eluted in 0.1 M glycine, pH 3.0 for 5 min at room temperature. Following addition of 1 M Tris, pH 8.5 for neutralization of samples, the beads were separated. The protein-containing supernatants were concentrated using Amicon Ultra 0.5 ml 3K centrifugal filters, separated on a NuPAGE 4-12% Bis-Tris Protein Gel and blotted. Membranes were probed with anti-FLAG antibody (F7425, SIGMA) and anti-HA antibody (H6908, SIGMA), both produced in rabbit.

Subcellular fractionation of C. crescentus NA1000

C. crescentus NA1000 cells expressing tagged MpcC were grown in PYE medium to an OD_{660} = 0.8 and fractionated as described previously with slight modifications (Anwari, 2012). The pellet from 5 ml culture was resuspended in 280 µl of ice-cold spheroplasting buffer (10 mM Tris, pH 7.5, 0.75 M sucrose). Lysozyme was added with final concentration of 100 µg/ml and cells were incubated on ice for 2 min. Following addition of cOmplete EDTA-free Protease Inhibitor Cocktail (Roche), 25 µg/ml DNaseI and 10 mM $MgCl_2$, two volumes of ice-cold lysis buffer (1.5 mM EDTA, pH 7.5) was added slowly to the cells with constant mixing. Cell lysis was completed using glass beads and unbroken cells were removed by two centrifugation steps at 4000 rpm for 10 min in a microcentrifuge. the supernatant containing the soluble fraction (periplasm and cytoplasm) was separated from the crude membrane fraction present in the pellet by ultracentrifugation at 50,000 rpm for 2 h at 4 ˚C using a TL100.3 rotor.

Cell membranes were fractionated according to (Thein et al., 2010). The inner membrane fraction was solubilized by resuspension of the membrane pellet in solubilization buffer containing 50 mM Tris, pH 8.0, 2% (w/v) Triton-X100, 10 mM

$MgCl_2$, and finally separated from the outer membrane fraction by centrifugation at 50,000 rpm for 2 h at 4 ˚C. The outer membrane pellet was washed in solubilization buffer followed by water and finally dissolved in Laemmli sample buffer. Proteins of the soluble and inner membrane fractions were precipitated with ice-cold acetone and solubilized in Laemmli sample buffer. Samples were analyzed by SDS-PAGE and western blotting using polyclonal anti-MpcC antiserum (1:1000, this work), anti-TimA (1:2000, this work), and anti-CpaC (1:1000, L. Shapiro, Stanford) produced in rabbit.

Proteinase K assay

C. crescentus NA1000 cells expressing HA-MpcC-$(FLAG)_3$ were grown in PYE medium to an $OD_{660} = 0.8$. Spheroblasts were prepared as described above. Following lysozyme incubation, Proteinase K was added to the cell suspension at a concentration of 0.1 mg/ml. Following incubation on ice for 30 min, Proteinase K was inactivated with addition of 10 mM PMSF. Spheroblasts were fractionated as described and analyzed by Western blot using rabbit anti-FLAG antibody (F7425, SIGMA) and anti-HA antibody (H6908, SIGMA).

Electron microscopy

For visualization of protein localization by Electron Microscopy **(**EM), *C. crescentus* NA1000 strain expressing HA-MpcC was cultivated in M2G medium to exponential phase. Protein expression was induced upon addition of 0.3% (w/v) xylose and cells were harvested 2 h after induction. Cells were washed twice in PBS and fixed in 0.1 M phosphate buffer, pH 7.4, containing 2% paraformaldehyde and 0.05% glutaraldehyde for 2 h at room temperature. Washed twice in 0.1% glycine, the samples were fixed and prepared for cryosectioning as described (Tokuyasu, 1973). Pellets were mixed with 10% warm gelatin and solidified on ice. Cut into blocks of about 1 mm^3, the solid mixtures were cryoprotected in 2.3 M sucrose at 4 °C overnight. The infiltrated blocks were frozen on cryosectioning stubs in liquid nitrogen and sections of 55–70 nm were cut using a Leica Ultracut UCT microtome equipped with a Reichert FCS cryo attachment. Retrieved with a 1:1 mixture of 2% methyl-cellulose and 2.3 M sucrose, the cryosections were placed on carbon/pioloform-coated EM support grids and floated upside down in PBS at 40 °C. For immunogold

labeling, sections were incubated with rabbit anti-HA antibody (Clontech), followed by incubation with secondary goat anti-rabbit IgG-ultra small gold antibody (Aurion). The sections were contrasted in methyl cellulose/uranyl acetate and analyzed on a FEI Tecnai G2 Spirit TEM (FEI, Hillsboro, Oregon, USA) operating at 120 kV. Images were taken with a Gatan Ultrascan 4000 (Pleasanton, CA, USA) camera at maximum resolution using manufacturer's software. MCUR1 fibril samples were directly applied to carbon/pioloform-coated EM support grids at a concentration of 0.05 mg/ml, washed with water and coated with uranyl acetate. After drying, grids were imaged as described.

NanoLC-MS/MS analysis

Coomassie-stained gel bands were digested in gel with trypsin as described previously (Borchert et al., 2010). Extracted peptides were desalted using C18 Stage tips and separated on an EasyLC nano-HPLC (Thermo Scientific) coupled to an LTQ Orbitrap XL (Thermo Scientific) as described elsewhere (Franz-Wachtel et al., 2012) with slight modifications: the peptide mixtures were injected onto the column in HPLC solvent A (0.1% formic acid) at a flow rate of 500 nl/min and subsequently eluted with a 57 min segmented gradient of 5-33-50-90% of HPLC solvent B (80% acetonitrile in 0.1% formic acid) at a flow rate of 200 nl/min. The ten most intense ions were sequentially isolated and fragmented in the linear ion trap using collision-induced dissociation (CID) at the ion accumulation target value of 5000 and default CID settings. Sequenced precursor masses were excluded from further selection for 90 s.

Acquired MS spectra were processed with MaxQuant software package version 1.5.2.8 with integrated Andromeda search engine (Cox and Mann, 2008; Cox et al., 2011). The database search was performed against a target-decoy *Homo sapiens* database obtained from Uniprot, containing 91,675 protein entries, the sequence of MCUR1-FLAG, and 285 commonly observed contaminants. No enzyme cleavage specificity was defined, and the minimum peptide length was set to five amino acid residues. Oxidation of methionine and N-terminal acetylation were specified as variable modifications, whereas carbamidomethylation on cysteine was set as a fixed modification. Initial precursor mass tolerance was set to 4.5 ppm (for the survey scan), and 0.5 Da for CID fragment ions. Peptide, protein and modification site

identifications were reported at a false discovery rate (FDR) of 0.01, estimated by the target/decoy approach (Elias and Gygi, 2007).

NMR spectroscopy

All spectra were recorded on Bruker AVIII-600 and AVIII-800 spectrometers. ^{15}N HSQC spectra were acquired over a temperature range from 298 K to 313 K. Diffusion ordered spectroscopy (DOSY) experiments were acquired to assess the translational diffusion times and obtain estimates of the effective molecular weight. Choosing the lower temperature, standard triple resonance experiments were acquired to perform backbone sequential assignment and 3D TOCSY spectra for sidechain assignment. These yielded assignments for a set of well-dispersed signals, consisting of a stretch of residues toward the N-terminus of the construct (C173-S190). However, few additional backbone assignments could be obtained, leading to considerable ambiguity and missing sidechain assignments, which made further analysis impractical.

Crystallization, data collection and structure determination

Crystallization trials were set up in 96-well sitting-drop plates with drops consisting of 300 nl protein solution and 300 nl reservoir solution (RS), and reservoirs containing 50 µl RS. Crystals of selenomethionine-labeled Kcr-0859ΔTM were obtained with a RS containing 100 mM sodium cacodylate, pH 6.5, 30% (v/v) MPD and 5% (w/v) PEG 2000. Crystals of CCDC90B$_{43\text{-}125}$-GCN4$_{N16V}$ were obtained using the Morpheus HT-96 screen (Molecular Dimensions), well F2. Prior to mounting, crystals of Kcr-0859ΔTM were transferred into a droplet of RS supplemented with 20% (v/v) glycerol for cryo-protection. All crystals were loop mounted and flash-cooled in liquid nitrogen. Data were collected at 100 K and a wavelength of either 1.07 Å (CCDC90B$_{43\text{-}125}$-GCN4$_{N16V}$) or 0.979 Å at the Selenium K-edge (Kcr-0859ΔTM) at beamline X10SA of the Swiss Light Source (Villigen, Switzerland), using a PILATUS 6M-F hybrid pixel detector (Dectris Ltd.). All data were indexed, integrated and scaled using XDS (Kabsch, 2010), with the statistics given in Table 1.

For the phasing of the Selenomethionine-labeled Kcr-0859ΔTM, we employed SHELXD (Sheldrick, 2008) for heavy atom location, locating six selenium sites in the

asymmetric unit. After phasing and density modification with SHELXE, one Kcr-0859ΔTM trimer could be traced with Buccaneer (Cowtan, 2006). The structure of CCDC90B$_{43-125}$-GCN4$_{N16V}$ was solved by molecular replacement with MOLREP (Vagin and Teplyakov, 2000), using trimeric coiled-coil fragments of PDB entry 5APQ search models. The structure was completed using Buccaneer. Both structures were finalized in cycles of manual modeling with Coot (Emsley and Cowtan, 2004), and refinement with REFMAC5 (Murshudov et al., 1999). Refinement statistics are given in Table 1, together with PDB accession codes.

Modeling

The comparative models of MCUR1 (residues 167-336) and full-length *Caulobacter* homologs YP_002517927 from *C. crescentus* NA1000 and WP_018113394 from *C. sp.* JGI 0001013-D04 were generated in Modeller (Sali et al., 1995) using the CCDC90B$_{43-125}$ crystal structure as a template for the head-neck segment. The backbone of the coiled-coil stalk was built using fragments of identical periodicities from solved crystal structures.

Figure 1

Click here to download Figure Fig1.tif

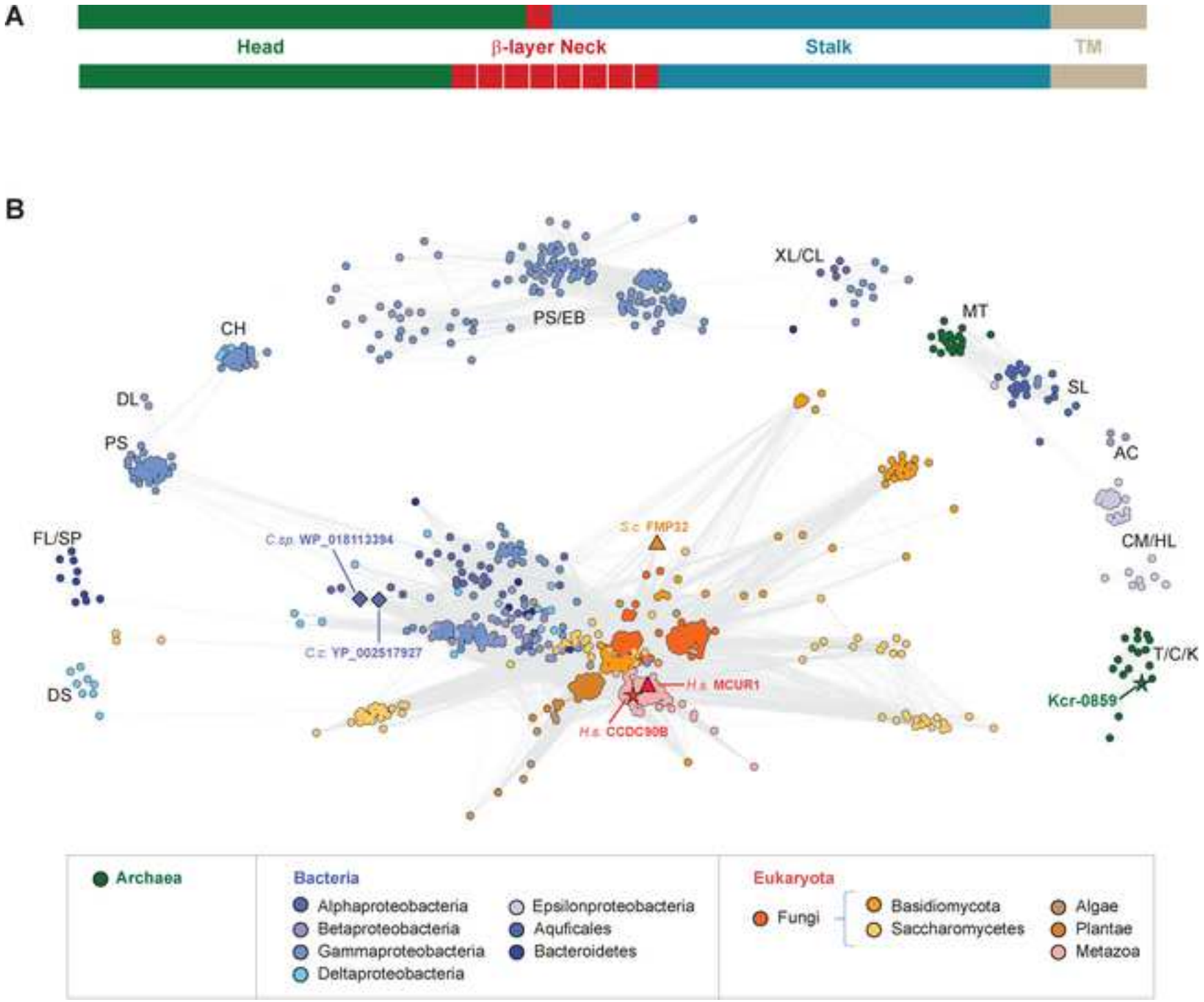


Figure 2

Click here to download Figure Fig2.tif

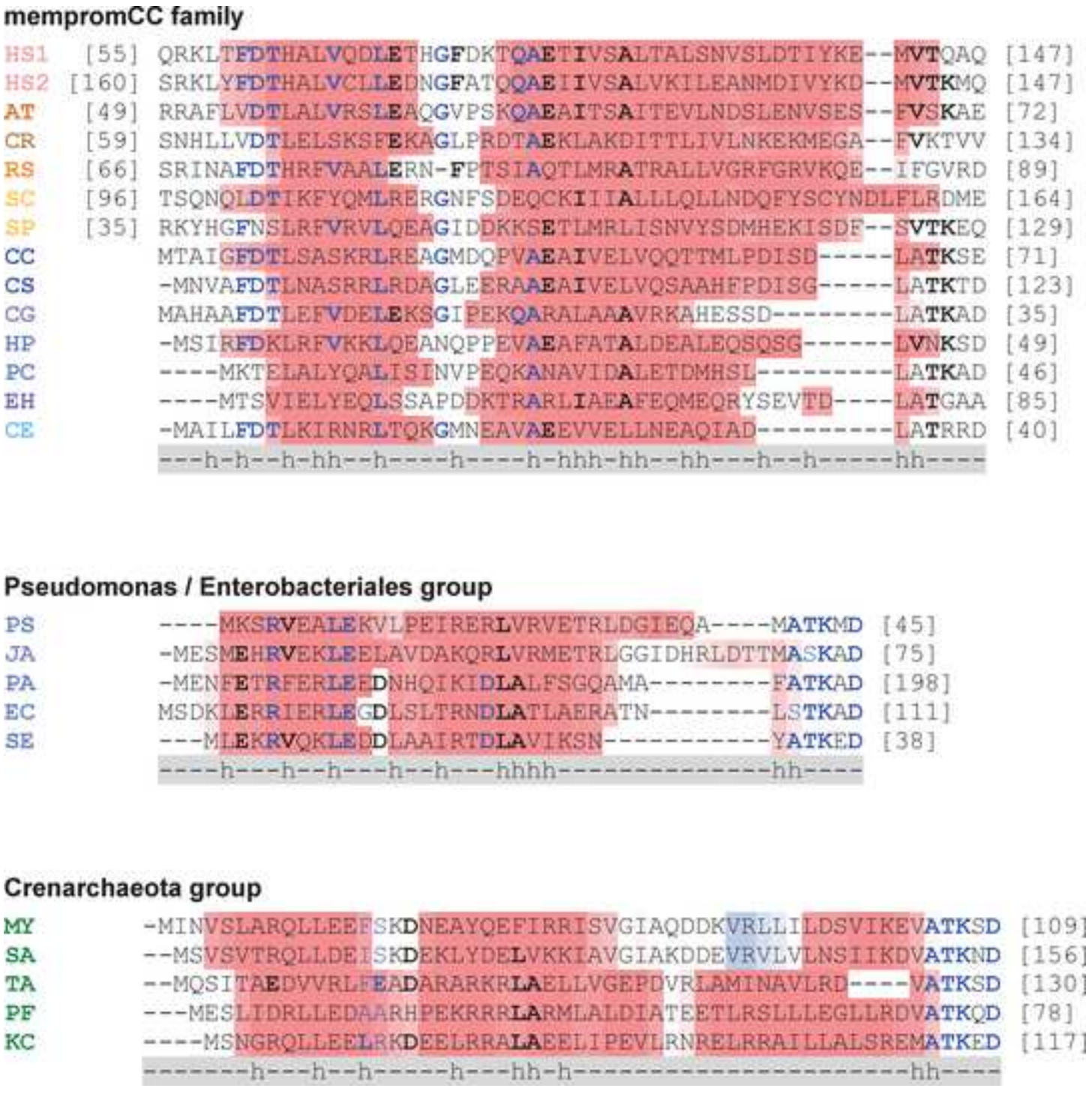


Figure 3

Click here to download Figure Fig3.tif

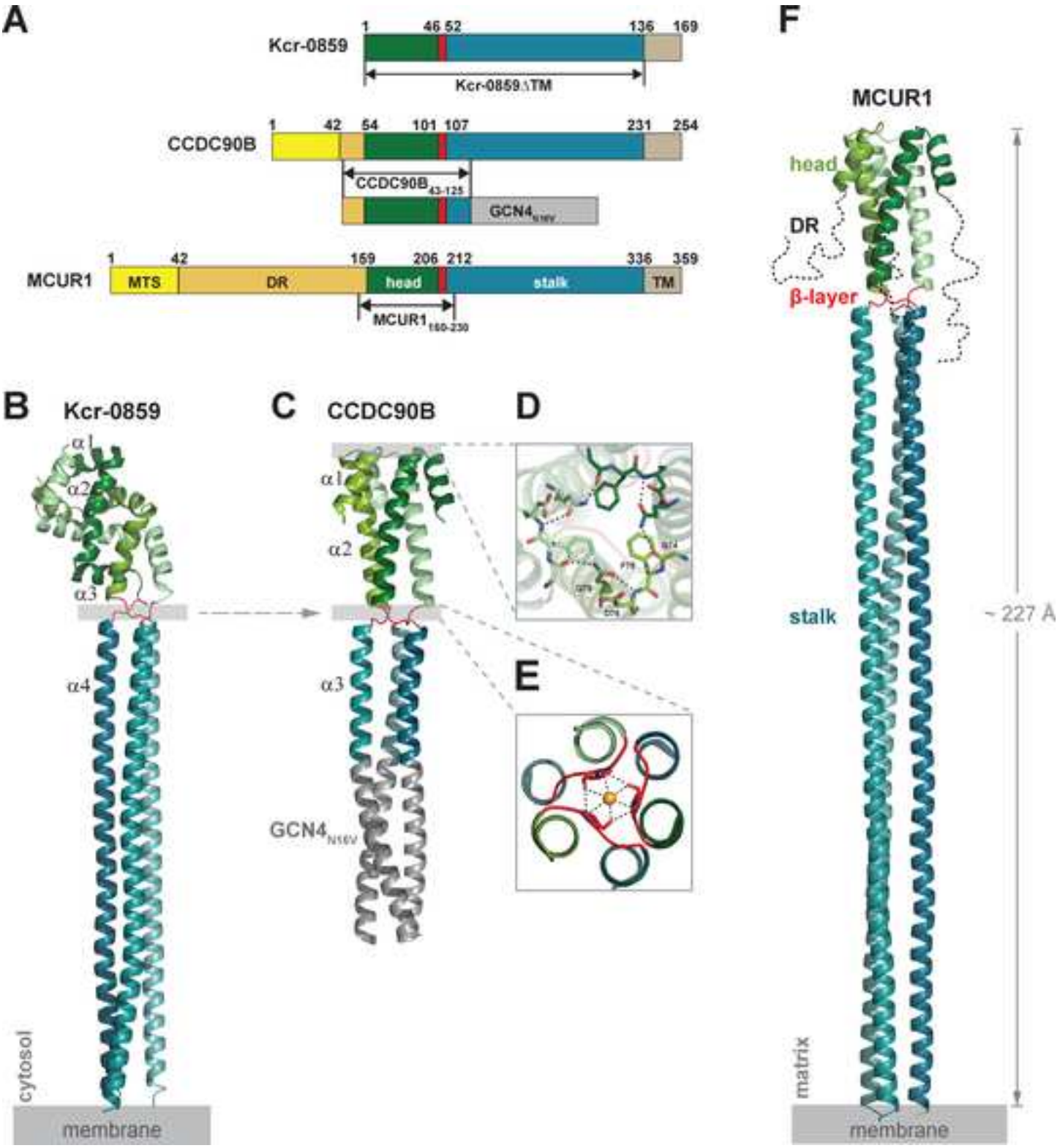


Figure 4

Click here to download Figure Fig4.tif

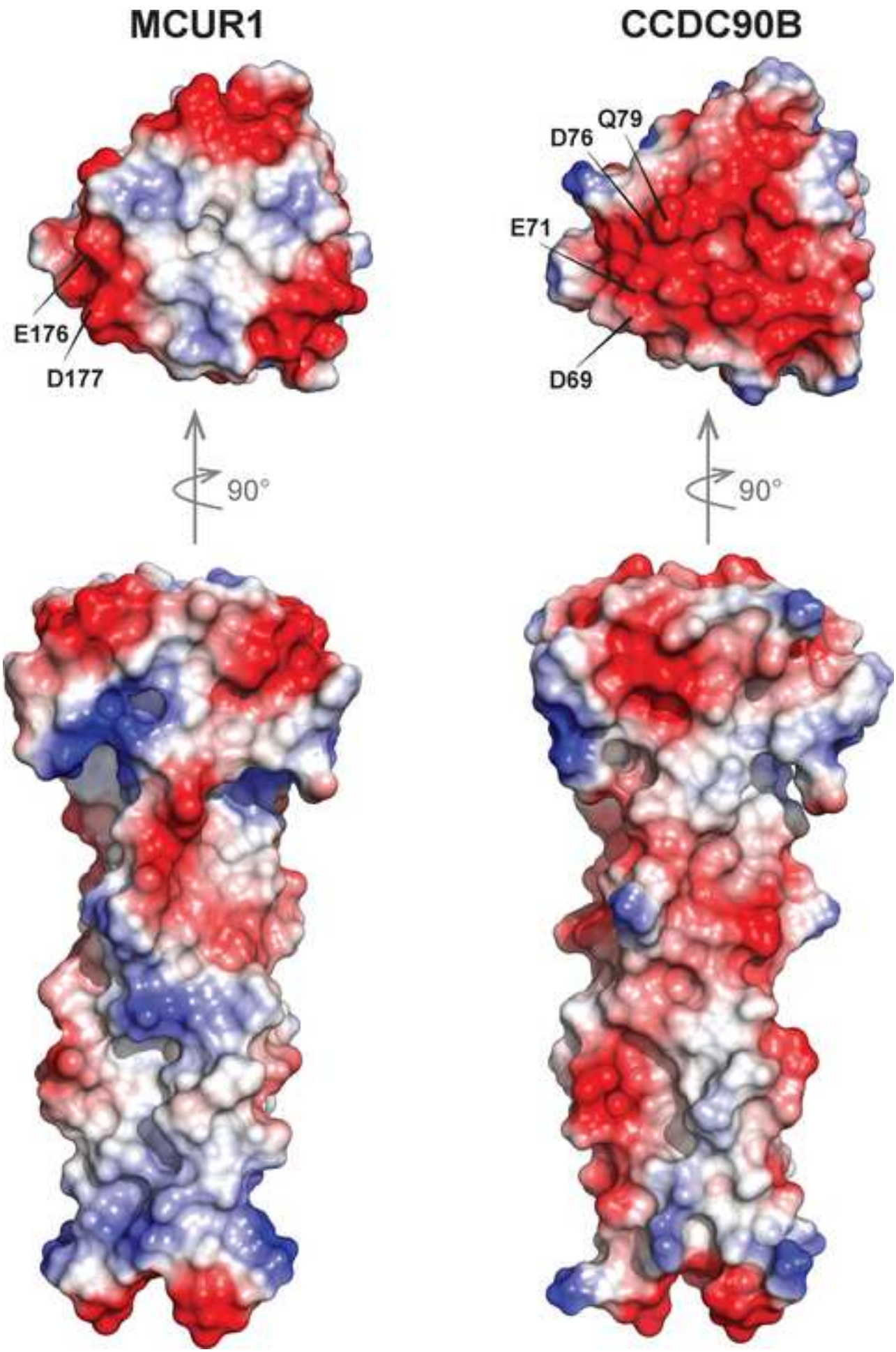


Figure 5

Click here to download Figure Fig5.tif

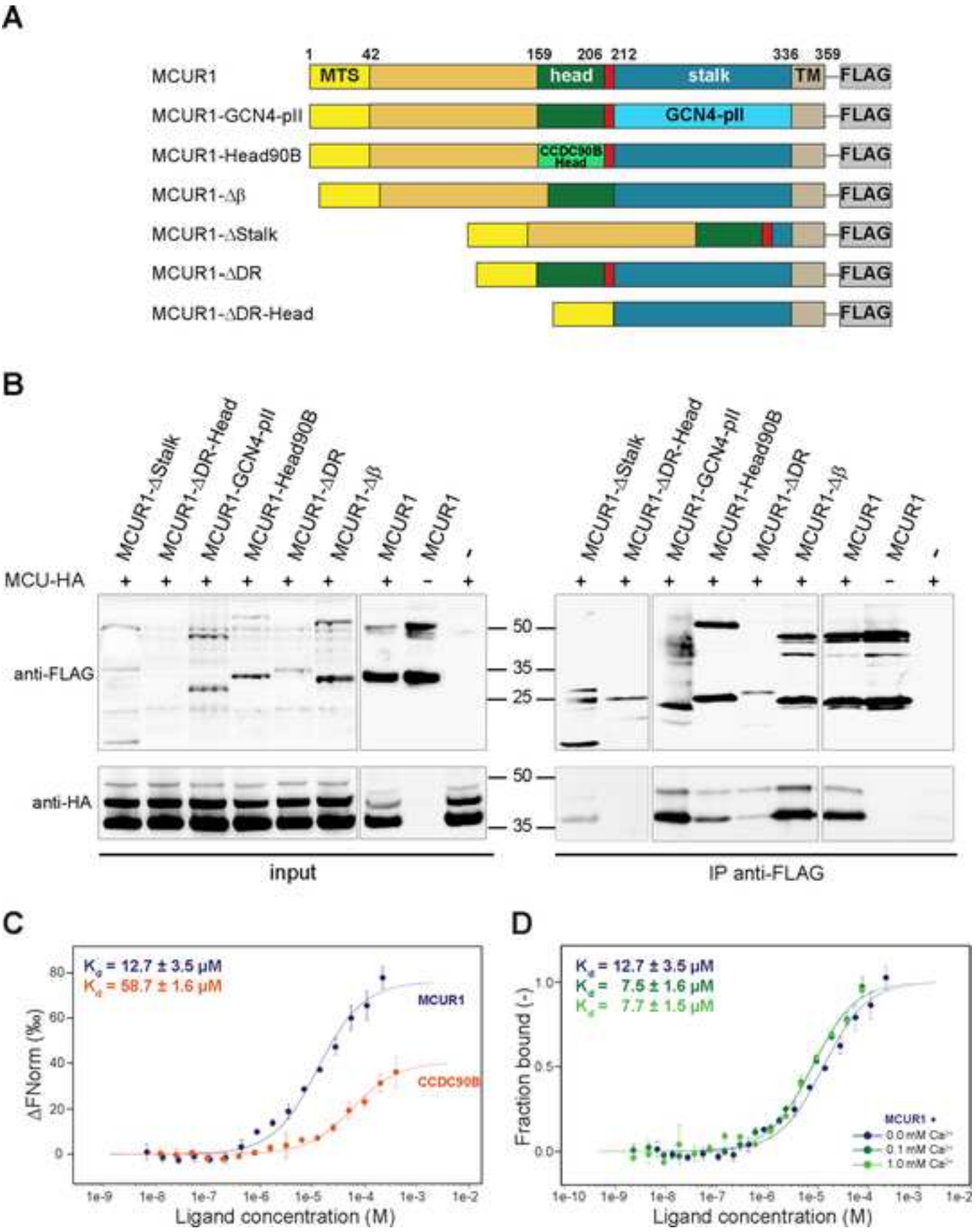


Figure 6

Click here to download Figure Fig6.tif

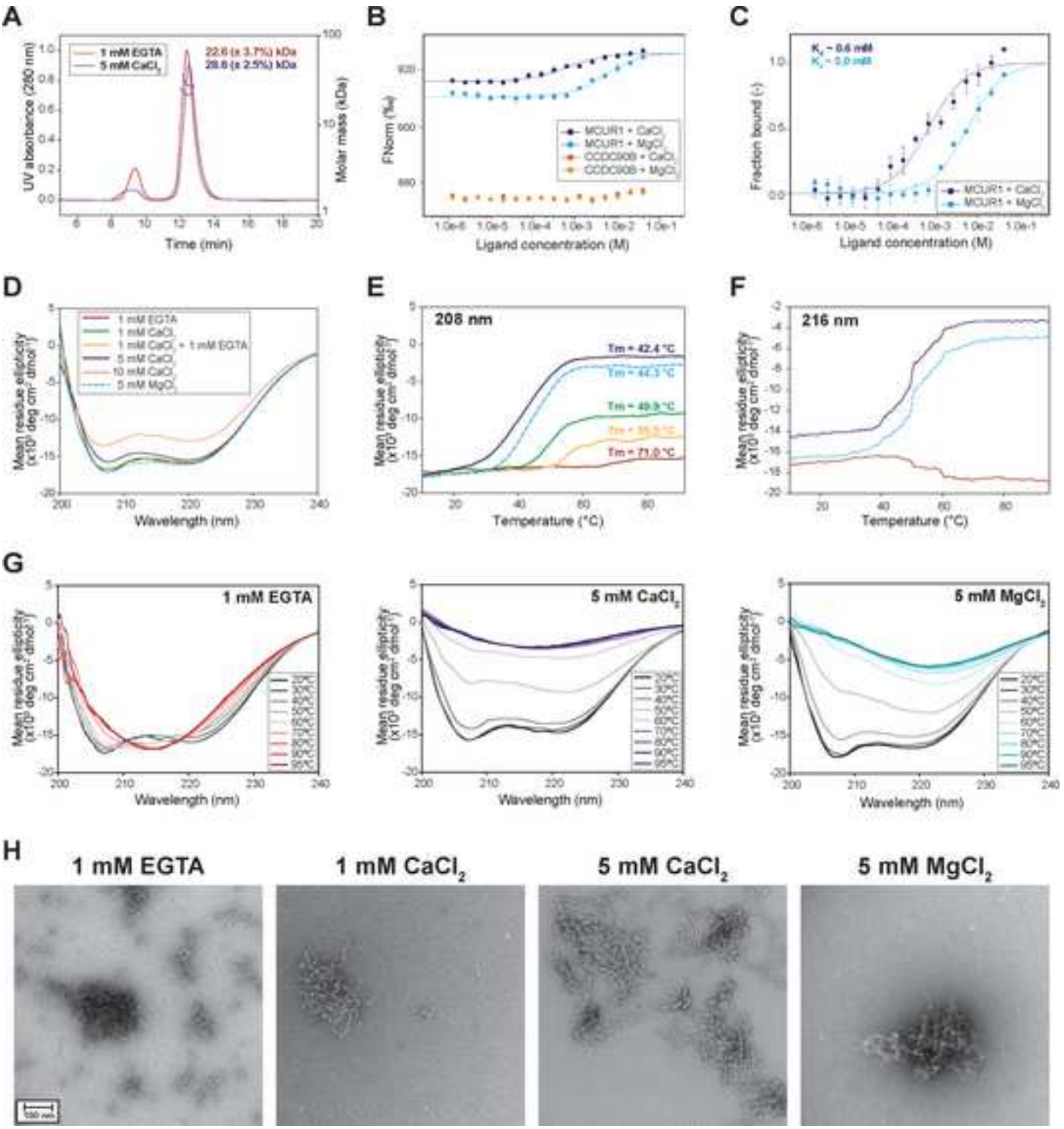


Figure 7

Click here to download Figure Fig7.tif

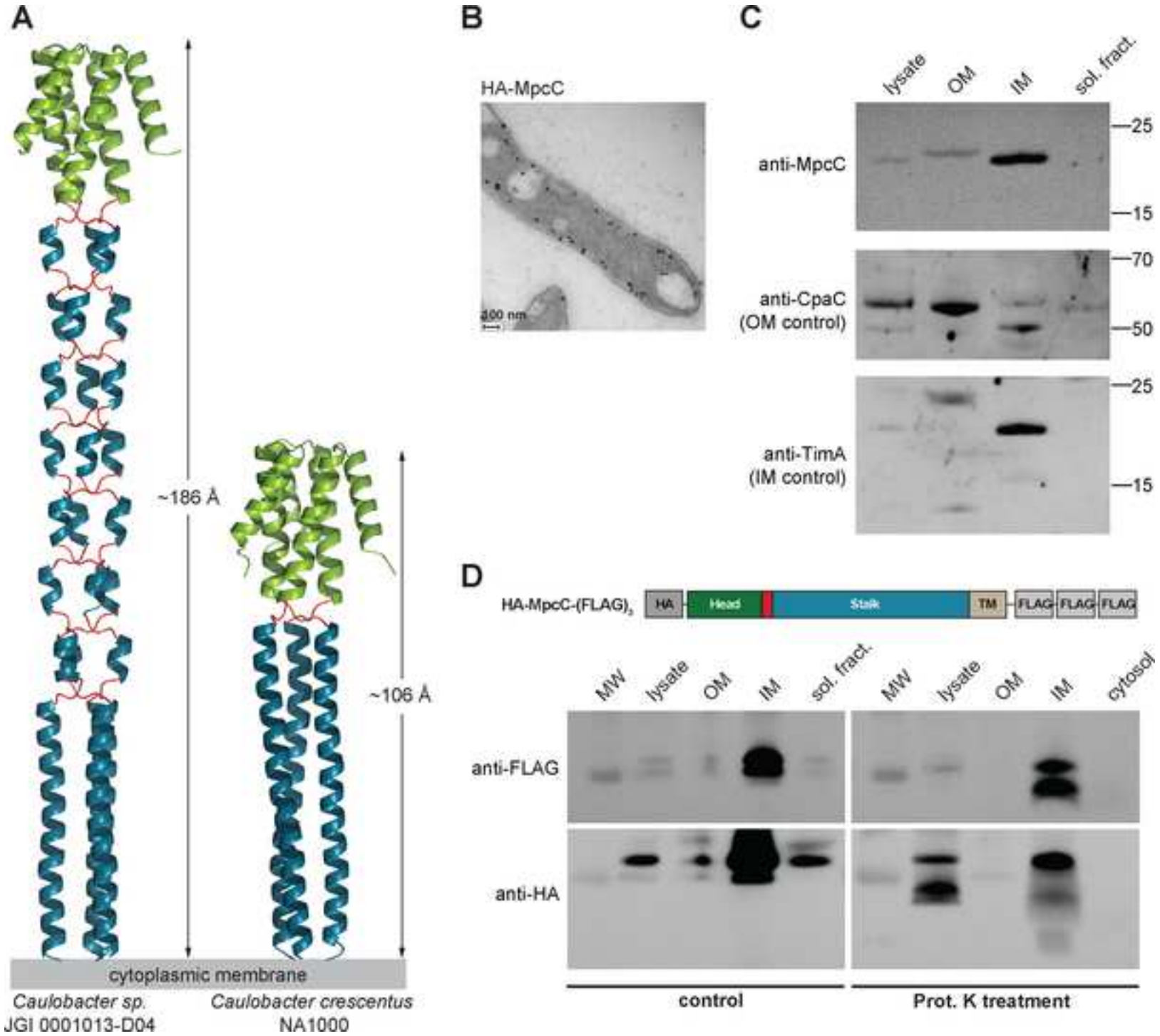
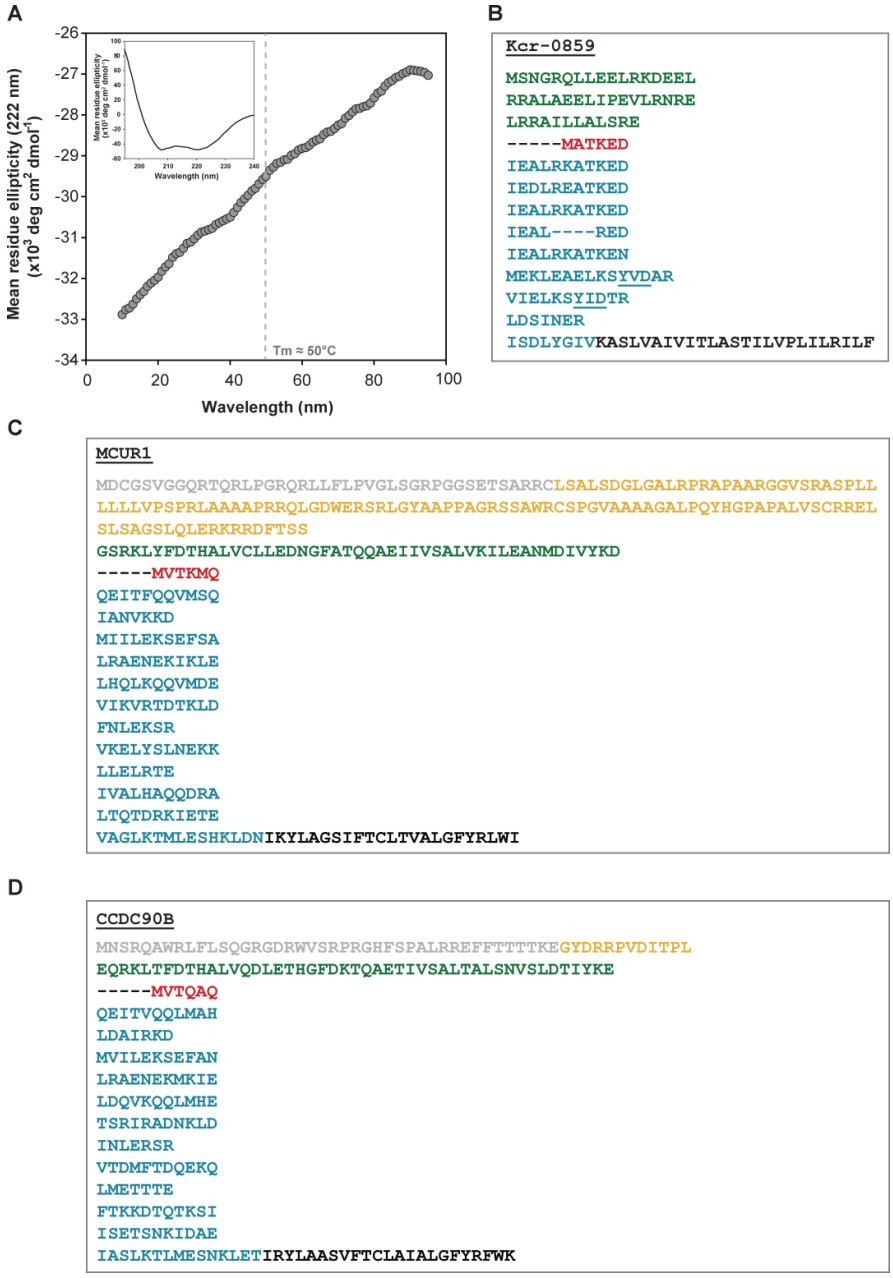
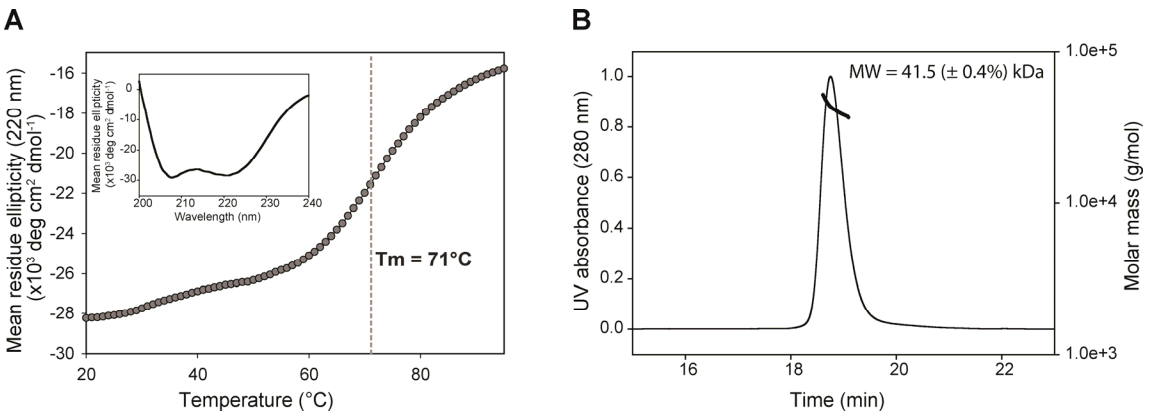


Figure S1, related to Figure 3.



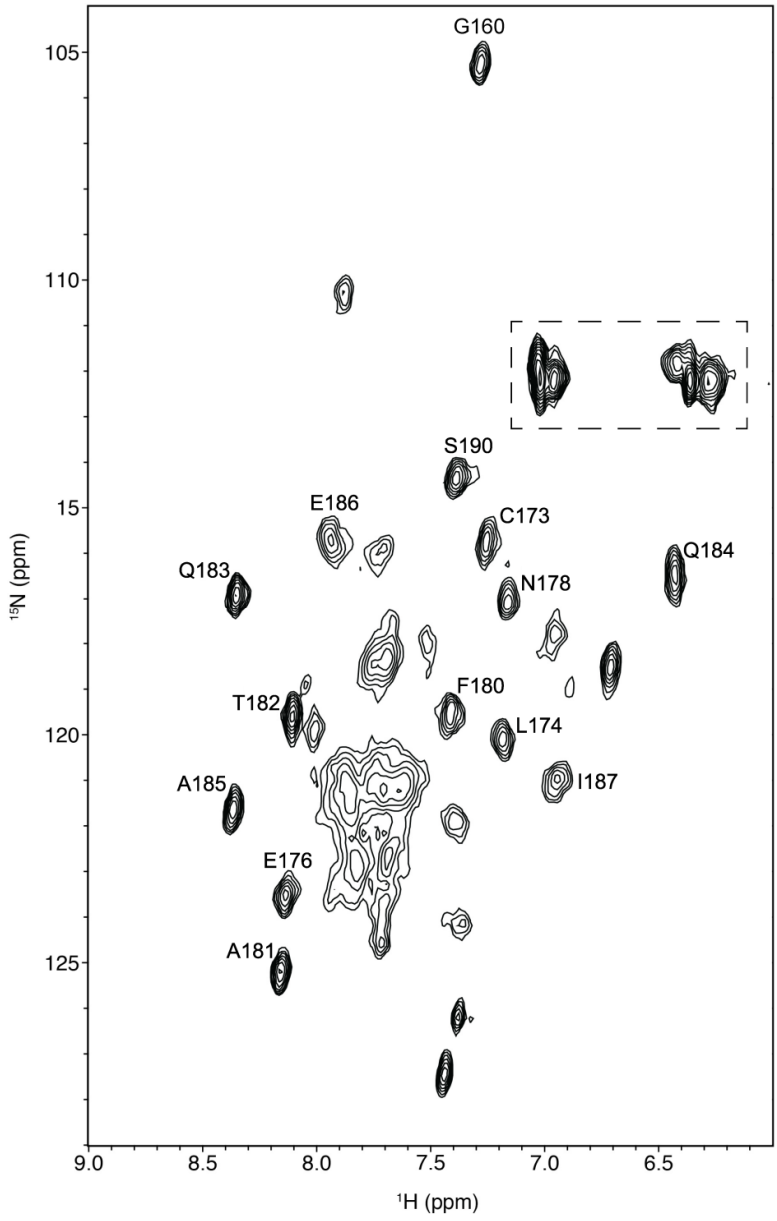
CD spectroscopy of Kcr-0859 and sequence annotation of Kcr-0859, CCDC90B and MCUR1. (A) Melting curve of Kcr-0859ΔTM monitored at 222 nm using CD spectroscopy. The protein melts with a melting temperature Tm ≈ 45°C. The inlet shows a single Far-UV CD spectrum measured at 20 °C. Sequence of crenarchaeal Kcr-0859 **(B)**, human MCUR1 **(C)** and CCDC90B **(D)** with individually marked domains and segments including the mitochondrial targeting signal (grey), the disordered region (orange), the head domain (green), the six-residue repeat forming the β-layer neck (red), the stalk (blue) and the helical transmembrane region (black). Polar YxD motifs in (B) are underlined.

Figure S2, related to Figure 3.



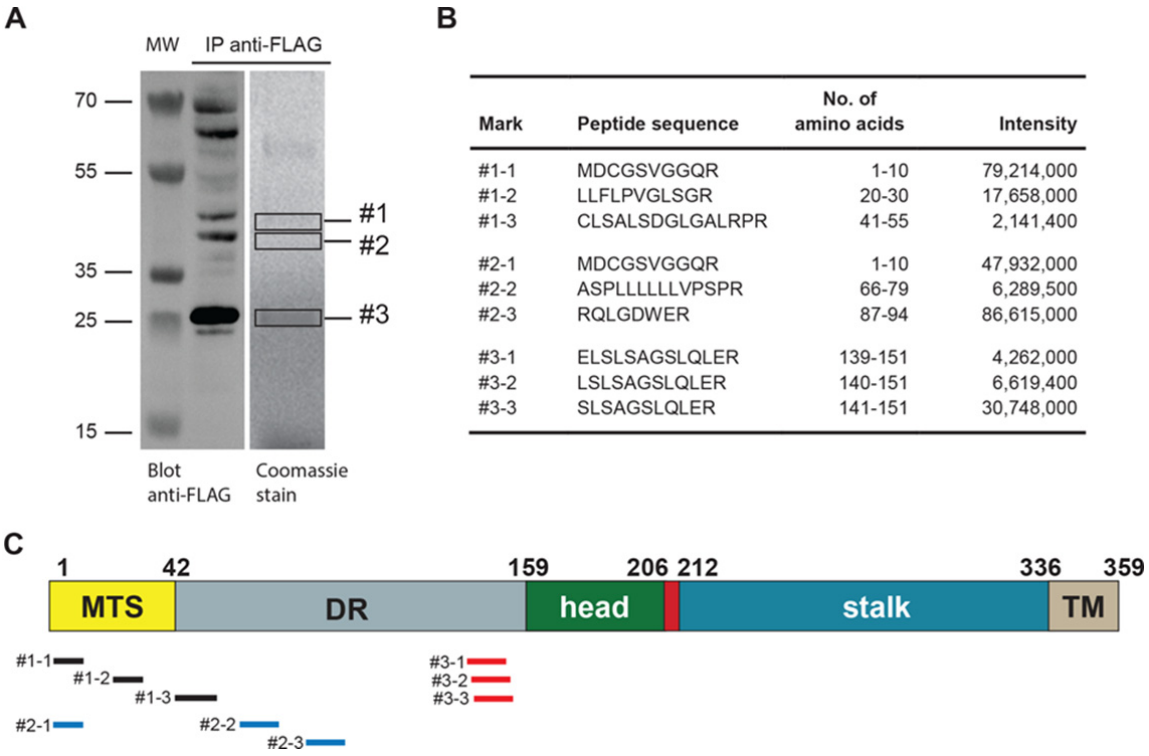
Biophysical characterization of CCDC90B$_{43\text{-}125}$-GCN4$_{N16V}$. (A) Melting curve of CCDC90B$_{43\text{-}125}$-GCN4$_{N16V}$ monitored at 220 nm using CD spectroscopy. The CCDC90B$_{43\text{-}125}$ fragment unfolds with a melting temperature Tm = 71 °C, thermal unfolding of the GCN4$_{N16V}$ segment with Tm > 95 °C is not visible in the applied temperature range. The inlet shows a single far-UV CD spectrum measured at 20 °C. **(B)** SEC MALS data of CCDC90B$_{43\text{-}125}$-GCN4$_{N16V}$ in 20 mM Tris, pH 7.5, 150 mM NaCl. The calculated molecular mass of the trimer is indicated.

Figure S3, related to Figure 3.



15N HSQC spectrum of MCUR1$_{160\text{-}230}$. Assigned residues are labelled. These consist largely of a stretch of residue between C173 and S190. Other peaks could not be unambiguously assigned. The boxed signals are due to side chain NH_2 groups of asparagine and glutamine residues.

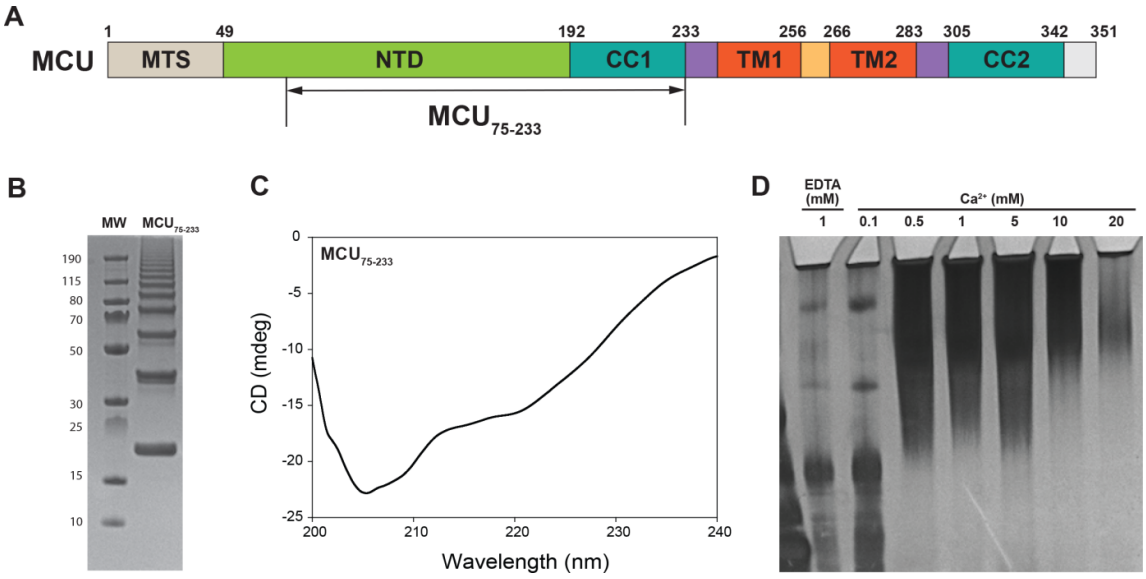
Figure S4, related to Figure 5



Mark	Peptide sequence	No. of amino acids	Intensity
#1-1	MDCGSVGGQR	1-10	79,214,000
#1-2	LLFLPVGLSGR	20-30	17,658,000
#1-3	CLSALSDGLGALRPR	41-55	2,141,400
#2-1	MDCGSVGGQR	1-10	47,932,000
#2-2	ASPLLLLLLVPSPR	66-79	6,289,500
#2-3	RQLGDWER	87-94	86,615,000
#3-1	ELSLSAGSLQLER	139-151	4,262,000
#3-2	LSLSAGSLQLER	140-151	6,619,400
#3-3	SLSAGSLQLER	141-151	30,748,000

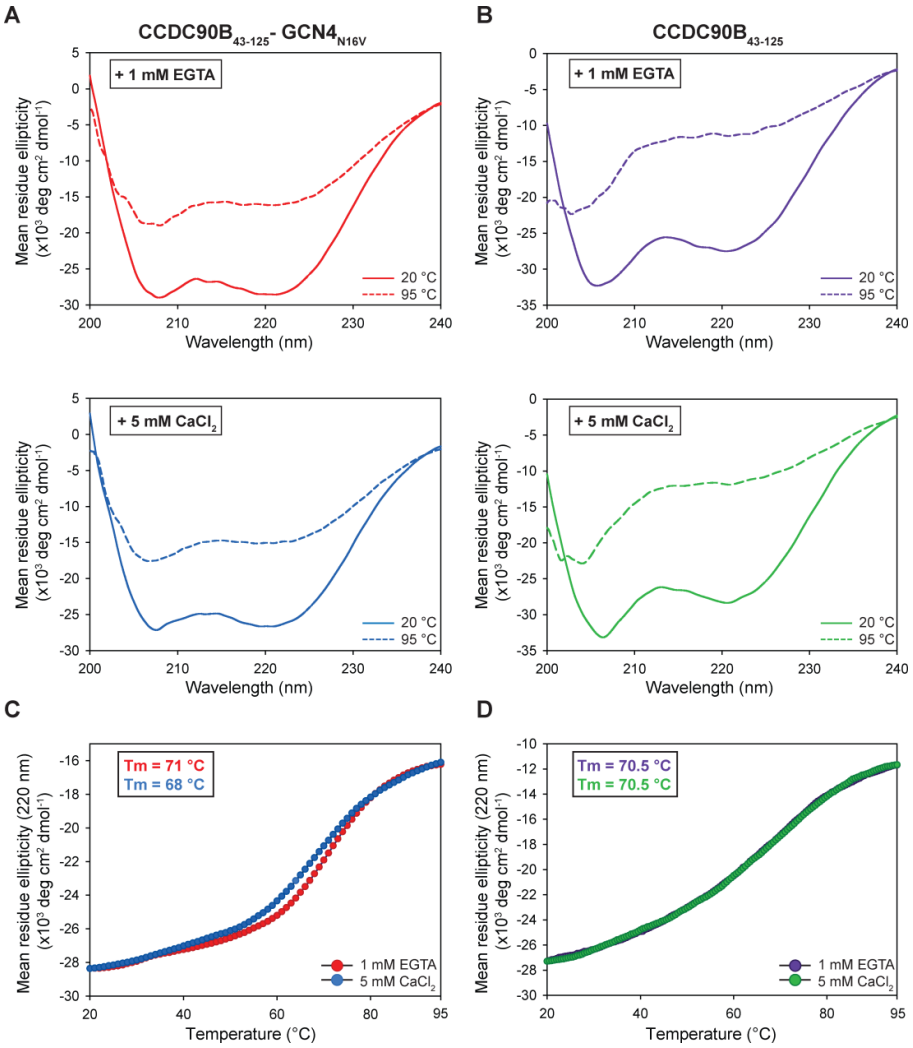
Mass spectrometry analysis of MCUR1 fragments. (A) Lysate of HEK293 cells, transiently transfected with MCUR1-FLAG, was subjected to immunoprecipitation using anti-FLAG antibody. A small sample was analyzed by western blotting using anti-FLAG antibody. The remaining sample was separated by SDS-PAGE and stained with Coomassie Brilliant Blue. Comparing western blot signals and protein bands of the stained SDS gel, the corresponding regions containing differentially processed fragments of MCUR1-FLAG were cut from the gel (#1, #2 and #3) and analyzed by mass spectrometry. **(B)** The table summarizes the sequence, residue numbers and intensity values determined by mass spec for the three most N-terminal fragments identified from each gel slice. **(C)** Schematic presentation of identified peptides in relation to the domain architecture of MCUR1.

Figure S5, related to Figure 5 and Figure 6



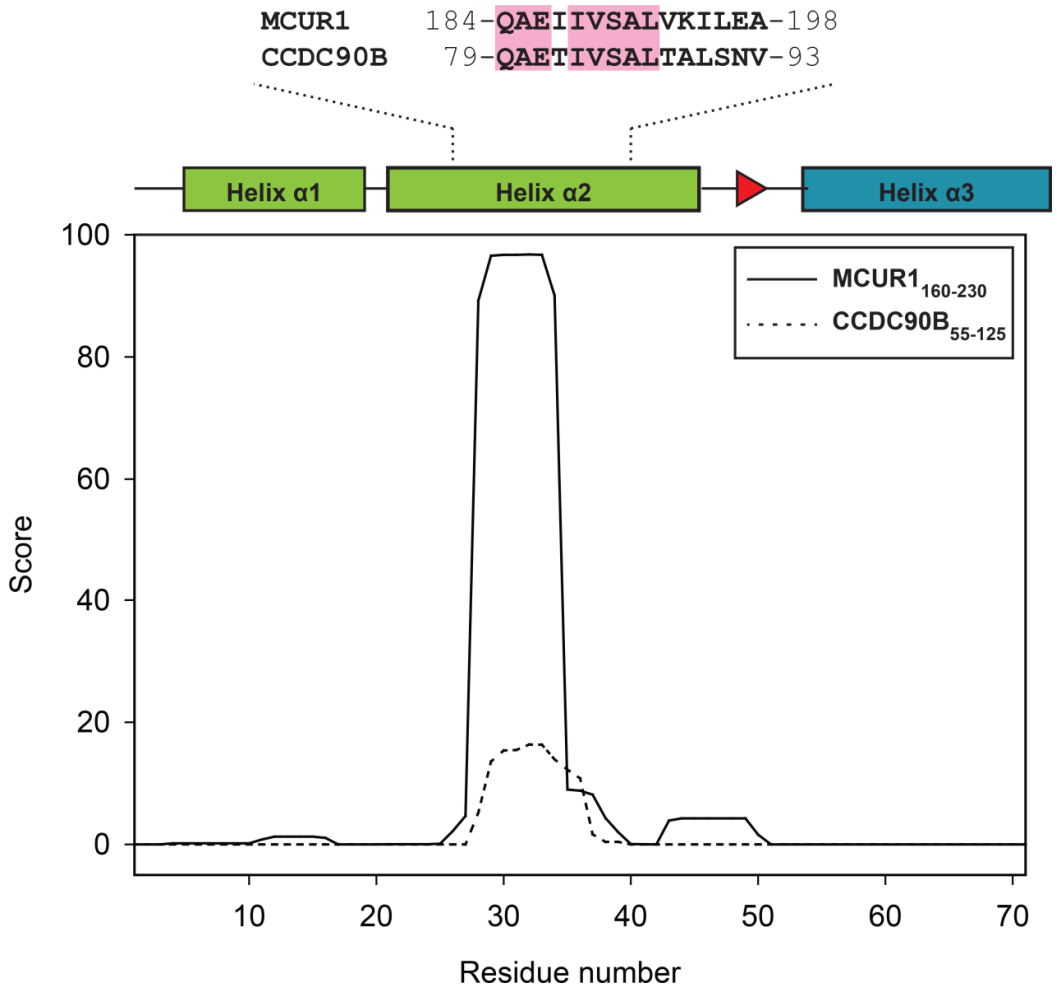
Oligomer formation of MCU_{75-233}. (A) Domain organization of human MCU (NP_612366.1) comprising of a mitochondrial target signal (MTS), an N-terminal soluble matrix domain (NTD), two transmembrane helices (TM1 and TM2), and two coiled coil domains (CC1 and CC2). As indicated, MCU_{75-233} includes residues 75-193 of the NTD succeeded by the CC1 domain. **(B)** SDS-PAGE analysis of MCU_{75-233}, which was purified under denaturing conditions and refolded. The protein forms a protein ladder of stable oligomers representing monomers, dimers, trimers, tetramers etc. according to their molecular weight in the gel. **(C)** Far-UV CD spectrum of MCU_{75-233} measured at 20 ˚C. **(D)** Blue-native PAGE analysis of oligomer and aggregate formation of MCU_{75-233} in dependence of increasing concentrations of $CaCl_2$. Aggregation was observed to start at concentrations of > 3 mM $CaCl_2$.

Figure S6, related to Figure 6.



Secondary structure analysis of CCDC90B$_{43-125}$-GCN4$_{N16V}$ and CCDC90B$_{43-125}$. Far-UV spectra of CCDC90B$_{43-125}$-GCN4$_{N16V}$ **(A)** and CCDC90B$_{43-125}$ **(B)** in the absence (1 mM EGTA) and presence of Ca(2+) (5 mM $CaCl_2$) recorded at 20 °C and 95 °C. Panels **(C)** and **(D)** show melting curves of CCDC90B$_{43-125}$-GCN4$_{N16V}$ and CCDC90B$_{43-125}$, respectively, recorded for each in the presence of 1 mM EGTA and 5 mM $CaCl_2$. Calculated melting temperatures (Tm) for each condition are indicated.

Figure S7, related to Figure 6.



Prediction of aggregation and amyloid propensity for MCUR1. The prediction tool TANGO was used to analyze the β-amyloidogenic propensity of $MCUR1_{160-230}$ and $CCDC90B_{55-125}$. The diagram shows the calculated TANGO scores for each protein fragment plotted against the residue numbers. The upper panel aligns the identified amylogenic hotspot of MCUR1 with the corresponding region of CCDC90B in relation to secondary structure prediction data.

Simulating primordial enzyme evolution: The creation of a bifunctional histidine kinase precursor by combining a *de novo* evolved ATP binding protein with the DHp domain found in the EnvZ histidine kinase.

Ioanna Karamichali[1], Edgardo[1,2] Sepulveda, Hongbo Zhu[1], Murray Coles[1], Andrei N. Lupas[1]*

[1]Department of Protein Evolution, Max-Planck-Institut für Entwicklungsbiologie, Tübingen, Germany, Max-Planck-Ring 5, 72076 Tübingen, Germany

[2]Present address: CONACYT, Department of Microbiology, Centro de Investigación Científica y de Educación Superior de Ensenada, Carretera Ensenada-Tijuana No. 3918 , Ensenada, Mexico

* correspondence to
Andrei N. Lupas,
Tel. –49 7071 601, 341
Fax –49 7071 601 349,
andrei.lupas@tuebingen.mpg.de

Key words: protein evolution, primordial enzymes, histidine kinase, ATPase

Abstract

We recreate the evolution process of a primordial enzyme by adding the natural dimerization and histidine phosphorylation (DHp) domain to an *in vitro* evolved ATP binding protein (DX), creating a fusion protein that simulates the cytosolic EnvZ histidine kinase of *E. coli*. The fusion protein successfully binds and hydrolyzes ATP *in vitro*, while significantly effecting the ompC promoter mediated gene expression *in vivo*. This is the first time a *de-novo* evolved protein has replaced a natural domain within the context of a whole functional protein system *in vivo*. This finding provides a proof of concept of the evolution of primordial enzymes, as well as the evolution of histidine kinases. Our approach can be further applied for the creation of novel enzymes and the facilitation of novel enzyme design.

Introduction

An increasing number of studies have revealed the RNA past of our world, where RNA was both storing genetic information and catalyzing chemical reactions[1–3]. However, the formation of the first cells would have been next to impossible without storing the genetic information on a more stable and reliable molecule like DNA, while chemical catalysis was shouldered by the more flexible and dynamic protein molecules. The critical step of the later transition came with the development of the ability of RNA to bind amino acids and to build up peptide bonds, a function it still retains until today as the main catalyst in the ribosome[4–8].

The first random polypeptides started out as simple protein folds which later formed protein domains characterized by crude, unspecific multifunctionality. The domains bound randomly to ligands and to each other, collaborating into functional novelty that could be stored in a newly formed genome as domain fusions, creating the first primordial enzymes[9,10]. In general the combination of domains leads to more sophisticated catalytic machines, allowing the selection towards more specific function while retaining the nature of a generalist in each single domain[11–13]. Multifuncionality allows enzymes to specialize towards different functionality[14–16] while domain reshuffling redistributes the functional vocabulary[17,18]. Primordial enzymes, like modern enzymes, could show a completely different or related functionality to their descendants, which could be performing only a part of the final catalytic process[14].

An example of domain evolution is presented by the catalytic ATP binding domain (CA) of histidine kinases that prevail in bacteria as the main environmental sensing system[19]. Both types of histidine kinases share the same CA domain while they greatly differ in their domain organization[20]. The combined DHp-CA domains form the catalytic part of the predominant type I histidine kinases, known as transmitter. The autophosphorylation of the histidine takes place within the transmitter, leading to the phosphorylation of an aspartate residue on the response regulator, which controls gene expression. The transmitter also functions as a phosphatase of the

phosphorylated OmpR, a feature specific to the DHp domain, which seems to be able to independently dephosphorylate OmpR[21]. During evolution more domains were gradually added to the transmitter, including the sensor domain which connects to the external environment, allowing the system to sense a variety of signals[20,22,23] .

One such sensor is linked through a HAMP domain (Histidine kinases, Adenylate cyclases, Methyl-accepting proteins and Phosphatases domain) to the transmitter of the extensively studied two components signal transduction (TCST) system EnvZ. EnvZ regulates the expression of membrane-pore forming proteins (porines) managing mainly an osmotic or even a chemical stress[24–26]. Sensor domains can vary greatly, indicating that different sensors have fused in the TCST systems during evolution[20]. This plasticity has allowed the creation of chimeras where the osmosensor of EnvZ has been exchanged with other sensors like the aspartate sensor Tar[27] (known as Taz TCST system) and the photosensor Cph1[28]. These chimeras are used to study TCST *systems in vivo* by coupling the chimeric histidine kinases to the expression of the green fluorescence protein (GFP) via the ompC promoter that is triggered from the phosphorylated response regulator of the EnvZ system, known as OmpR[29–31].

The CA domain reveals an interesting evolutionary fact: Primarily, it is highly conserved in both types of histidine kinases, but homologs have also been found in all members of the GHKL (bacterial gyrase, HSP90, histidine kinase, MutL) superfamily[32–34] and even as independent domains in homodimeric serine kinases[35–37]. These findings lead to the hypothesis that the original purpose of the CA domain was to bind ATP and later was combined with different domains participating in more complicated functions like ATP hydrolysis, and phosphorylation of either histidine or serine amino acids. In contrast DHp was only found in combination with the CA domain, indicating an early in evolution fusion of the two domains.

The homology of the CA domain is based on the detection of the highly conserved N,G1,F,G2 amino acid boxes and also a shared structural architecture of a 4 α-helices/5 β-strands sandwich, known as the Bergerat fold[20,38,39]. Most differences between the homologs can be found on different loops of the molecule while some homologs have extra β-sheets, or α-helices

like EnvZ[20]. The ATP binding pocket is found between the β-sheets β5-β7 and the helices α2, α3 and α5, bordered by the four conserved boxes and further sealed by a flexible long loop as shown in the cartoon representation of the EnvZ CA domain in figure 1.A.

A simpler ATP binding protein was *de novo* evolved by Keefe and Szostak (2001). This novel protein called DX was created *in vitro* through multiple cycles of mutagenesis and selection, based on the ability to bind ATP[40]. DX is a zinc finger protein that forms an α-helix-three β-sheets-α-helix fold. The ATP binding site is a hydrophobic pocket found between the three β-sheets and α-helix α2, which are connected by a loop that borders the binding site. The binding pocket is rather shallow leaving exposed the sugar and phosphate moieties of the ATP molecule (Figure 1.B.)[41,42]

In this paper we recreate the evolution of a primordial enzyme precursor of a histidine kinase by fusing DX with the histidine-containing domain DHp which is responsible for the dimerization and also the binding and phosphorylation of the OmpR response regulator. This fusion led to a bifunctional protein which successfully binds and hydrolyzes ATP *in vitro*, while significantly effecting the ompC promoter mediated gene expression *in vivo*. It is the first time that a *de novo* evolved protein was successfully used to replace a natural domain within a functional system *in vivo*, while providing a proof of concept for the evolution of primordial enzymes. Our approach can be used to further aid the challenging enzyme designing process, through the combination of natural domain recycling, *de novo* protein evolution and *in vivo* selection for the creation of novel enzymes.

Methods

Protein design

Three fusion models of the published structures of DHp and DX[43–45] where manually designed facilitating the close proximity of the bound on the DX domain-ATP and the histidine residue of the DHp domain (around 4Å). The accuracy of the models DX-DHp, DHp-DX1, and DHp-DX2 was tested using the Fast Relax protocol from the Rosetta software (40000 repeats)[46]. In both DHp-DX1 and 2 linkers were designed between the DHp and DX domains. DHp-DX1 has the natural linker found in the EnvZ histidine kinase while DHp-DX2 has a glycin-rich linker designed using the Rosetta Loop Design protocol[24]. No linker was used for DX-DHp where the domain order was inversed in comparison to the natural DHp-CA transmitter.

Construct design

The genes of the designed proteins were synthesized based on the amino acid sequences of the DHp-DX1,2 and DX-DHp models (Eurofins Genomics). The vectors pCL1920(taz) and pET3a(gfp(LVA)[27,29,47], facilitated the design of the reporting system for each fusion along with the controls used (DHp, DHp-CA). A His^Gln mutant DHp-DX1(H15Q) was constructed through round the horn amplification for each one of the pCL1920(DHp-DX1) and pET-TRX(DHp-DX1) constructs, targeting the histidine in positions 23 and 151, respectively. We used PCR to amplify the response regulator OmpR and the protein DX. All genes were modified to include the flanking restriction sites HindIII/BamHI and/or NcoI/BamHI, which facilitated the cloning into the vectors pCL1920 and pet-TRX, respectively. All TRX fused proteins had an N-terminal 6xHis-tag sequence right after the TRX protein, allowing their further nickel affinity purification and anti-His tag antibody detection. All primers used are presented in the supplementary table 1 (supplementary material).

In vivo functional characterization

We used the cell line *E. coli* JM1012 (P_{ompC}::cI-LVA ΔenvZ) and plasmids pJM1 and 2, which differ only on the ribosomal binding site used (pJM1:strong, pJM2:weak[48]). pCL1920 constructs were cloned along with the pJM1 or pJM2 vectors in JM1012 cells. Cells from a frozen glycerol stock were grown in 5ml LB overnight at 30°C, 50μl of the culture were transferred into 5ml LB and they were grown at 37°C for 2 hours. The OD was adjusted to 0.3 (570nm), the cells were induced with 0.25mM IPTG and 10mM Asp and 200μl were placed in each well of a 96 well plate (black with clear bottom, from Sigma-Aldrich). The plates were incubated at 37°C overnight, while fluorescence measurements were taken every 30min with an excitation of 420 and an emission of 565nm using the Synergy H4 Hybrid Reader (BioTek Instruments). All experiments were performed in the presence of kanamycin (Roth) and spectinomycin (Sigma Aldrich) and were repeated in minimal media as well. Each sample was replicated 6 times, the experiment was repeated three times. The final results represent the average of these measurements after two hours incubation in LB.

Protein overexpression

BL21Gold *E.coli* cells transformed with the pETTRX vectors containing genes for either the DX-DHp, DHp-DX1 and 2 proteins were taken from frozen glycerol stocks and grown overnight in LB at 30°C. Cultures were induced with 0.25mM IPTG in the presence or absence of 2% ethanol or 0.2mM NaCl, at an OD of 0.6 at 30°C, for 4h. Cells were harvested by centrifugation at 6000rpm for 15 min and directly frozen at -80°C. Subsequently, the cells were thawed on ice, resuspended in lysis buffer [Buffer A: 30mM Tris, 500mM NaCl, pH 7-8.5) complemented with lysozyme (Sigma Aldrich), DNase I (AppliChem), and 1 Protease Inhibitor Tablet (cOmplete Roche)], and further lysed via sonication. Soluble and insoluble protein fractions were separated through centrifugation at 16000rpm for 60min. Protein expression and

solubility was tested in SDS PAGE (12%) using 4xSDS loading buffer [62.5mM Tris, 2% SDS, 10% glycerol, 5% 2-mercaptoethanol, 0.001% bromphenolblau, pH 6.8], and 15µl of the samples. 20µl of the final mixture were boiled and loaded in the gel. The run was of 30mA/gel for around an hour.

The proteins DHp-DX1 and 2 fused with TRX were more stable when overexpressed in LB, after a 4h induction with 0.25mM IPTG and 2% ethanol at an 0.6 OD, at 30°C. The same conditions were used for overexpression of DHp-DX1(H151Q).

The TRX fusion of DHp-CA was overexpressed the same way, but induced with 0.25mM IPTG at 37°C. The TRX fusion of OmpR was expressed in Rosetta 2 (DE3) pLysS cells. Cells were induced with 0.25mM IPTG in 16^{o}c for 16h. Finally, the TRX fusion of DX was expressed in Rosetta 2 (DE3) pLysS cells induced with 0.25mM IPTG and 0.2mM NaCl for 4h at 37^{o}C.

Protein purification

We purified the TRX fused DX-DHp, DHp1 and 2 proteins using nickel affinity chromatography in Buffer A and B (Buffer A plus 0.5M of imidazole). The proteins were dialyzed and concentrated in Buffer C (30mM Tris, 200mM NaCl, 5mM citric acid, 5mM 2-mercaptoethanol, 5µM $ZnCl_2$, 10% glycerol, pH 8.5).

The TRX fused DHp-CA, DX and OmpR proteins were purified identically and dialyzed in Buffer D (30mM Tris, 150mM NaCl, pH 7-8.5). 50% glycerol stocks were prepared for each protein and were frozen in liquid nitrogen to be stored at -80^{o}C as 1mg/ml aliquots. Additionally, the TRX protein was cleaved using TEV protease (dilution 1 to 40) at -4^{o}C, overnight. The proteins were further purified using Ni^{2+} affinity chromatography to remove the TRX fusion protein and the TEV protease. Finally, the proteins were dialyzed in Buffer E (0.1mM Tris, 50mM KCl, 5mM β-mercaptoethanol) frozen in liquid nitrogen and stored at -80^{o}C, in aliquots of 5% glycerol and 1mg/ml protein.

Biophysical analysis

Purified proteins were loaded on size exclusion columns and eluted using the selected storage buffer of each protein as mentioned above. The protein DHp-DX1 was further dialyzed in Buffer D and analyzed using Circular Dichroism (0.3mg/ml, pH 8.5, J-810 Spectropolarimeter JASCO) and one-dimension (1D) NMR (7mg/ml, pH 8.5).

N1 anti-phospho-histidine western blot

Both, self-prepared 15% polyacrylamide native gels and commercial (Invitrogen) 4-16% gradient native gels were used to immobilize the proteins in native conditions (pH 7-8.5), in the presence and absence of 0.002% w/v Coomassie R-250 in the cathode buffer (blue and clear native page). Gels were loaded with 15µl of protein sample and run for 1.5 hours at 300V (no more than 15mA).
Western blot was used for the detection of a potential histidine phosphorylation at the N1 position of the histidine. Proteins were transferred to Protrans 8A-PVDF membranes, and detected using the N1 anti-phospho-His antibody (Hölzel Diagnostika) and an anti-rabbit secondary antibody (Dianova: Goat anti-rabbit HRP). The results were visualized with the Clarity Western kit (Biorad).

Autoradiography

The TRX cleaved proteins in 5% glycerol stocks (see above) DHp-CA, DHp-DX1, DHp-DX1(H15Q) along with the TRX fused DX protein, were mixed with the OmpR response regulator at concentrations of 2 to 20µM. The samples were screened under addition of either $MgCl_2$, $CaCl_2$ and $MnCl_2$ (5-10mM) in the presence of a mixture of cold (50µM) and radiolabeled ATP (0.2-1µl of gamma-^{32}P ATP; 37MBq/ml). Buffer D was used, at pH 8, 28^{o}C and an incubation time within 60 minutes. After incubation, 4x loading buffer was

added in each sample. The samples were heated for 2 min at 95^{o}C and loaded on either an in-house (18% SDS PAGE, in Anode/Kathode buffer: 25mM Tris, 0.2M glycin, 0.1% SDS, pH 8.5) or a commercial SDS gel (12% Invitrogen, in running buffer: 50mM MOPS, 50mM Tris, 0.1% SDS, 1mM EDTA, pH 7.7). The gel was run at 25-30mA for 1.5h and finally washed, shelled and analyzed via autoradiography (24h exposure time, visualized by the Fuji Imager FLA 3000).

In gel ATPase activity test

15% polyacrylamide native gels were used for detection of ATPase activity using the high sensitivity lead-phosphate reaction[49]. After running the gel, it was further incubated in buffer 40mM Tris, 4mM ATP, 1.5mM $MgCl_2$, 1.5mM $ZnCl_2$, pH 8.5 for an hour. Subsequently, the gel was placed overnight in a lead phosphate reaction buffer (35mM Tris, 270mM glycine, 20% v/v methanol, 0.075% w/v lead nitrate, 0.8mM ATP, 14mM $MgSO_4$, 14mM $ZnSO_4$, pH 8.5). The gels were imaged the next day, after being exposed for 30sec in an ammonium sulfide solution (1% v/v ammonium sulfide) to develop the color of the band. Brown (lead sulfide) bands were expected on the gel where free phosphate was released. Three short washing steps with 35mM Tris in pH 8.5 followed each buffer exchange. The gel was finally cut into two; with both parts including identical samples, the first half was dyed using coomassie, while the second half was used for a semi-dry western blot (as above). In this case, anti-His Tag antibodies (Sigma) were employed for the verification of the position of the immobilized proteins of interest.

ATPase functional characterization

Both colorimetric (Phosphate Assay Kit, Abcam) and spectrophotometric (EnzChek™ Phosphate Assay Kit, Thermofisher) methods were employed for the characterization of the ATPase activity of DHp-DX1. First, the optimal conditions of the ATPase activity, along with the optimal protein concentration of use, were identified colorimetrically with the end point-malachite green phosphate assay kit (Abcam). The screening included protein concentrations of 2 to 20μM, in buffer D with pH of either 6, 8 or 9.5; and the addition of either $MgCl_2$, $CaCl_2$ or $MnCl_2$ (10mM), in the presence of 100μM ATP. Samples were incubated for 0, 15, 30 and 60min at 28 or 37^{o}C. The same screening was applied for the TRX fused DX protein which was used as a negative control.

The less sensitive continuous, spectrophotometric and MESG/PNP(2-amino-6-mercapto-7-methylpurine riboside/ purine nucleoside phosphorylase)-dependent EnzChek™ phosphate assay kit (Thermo Fisher Scientific) was used for a kinetic characterization of the ATPase activity. The analysis was performed using 2μM DHp-DX1 in buffer D, pH of 6 and 10mM $MgCl_2$. The ATP concentrations used varied between 50 and 1000μM. The same analysis was performed for ADP and GTP in order to assess the specificity of the enzymatic activity. The proteins used were cleaved from the TRX fusion using TEV protease (1 to 40 dilution) and stored in 5% glycerol as described above.

Results

Protein design

The three DHp, DX fusion models where manually designed for a close proximity between the histidine on the DHp domain and the ATP bound on the DX protein (Figure 2.A. and B.). The DHp-DX1 had a linker similar to the natural EnvZ linker (TGQEMPG) between the two domains, while DHp-DX2 had a designed glycine-rich linker instead (DKGGGGG). All models were relaxed using the FastRelax protocol to eliminate any molecular clashes and represent the most plausible conformation of the structure. All models can be found in the supplementary material (Supplementary Figure 1). In this study we mainly focus on the most successful of our fusion proteins, the DHp-DX1 protein (Figure 2.C.).

In vivo functional analysis

Even though the *in vivo* phosphorylation and dephosphorylation of the OmpR response regulator has been observed multiple times in the past[21,29,50], the cross-phosphorylation of OmpR[51,52] along with other weaknesses of the systems used (variation in vector copy number, leaky gene expression, high stability of the reporter protein) were increasing the background or causing a reduction in sensitivity. We here introduce a cell system that allows the *in vivo* detection of either a phosphorylation or dephosphorylation of OmpR*,* by using a minimized background as a baseline to distinguish between the two functionalities. Our TCST system is a combination between the Taz-chimeric system and the Cph1-chimeric system, introduced by Michalodimitrakis et al. (2005) and Lee et al. (2013), respectively.

We have replaced the Cph1-chimeric kinase with either the Taz-chimeric kinase, parts of the kinase (DHp-CA, DHp), or any of our DX fusion proteins carried by the low copy number vector pCL1920[29] in *E. coli* JM1012[30]. This combination allowed the simultaneous detection of both OmpR

phosphorylation and dephosphorylation, by using the remaining background signal as a point of reference for any induction or reduction of GFP fluorescence.

The function of the protein of interest alters the phosphorylation state of OmpR that in turn alters the expression of the CI-LVA repressor via the OmpC promoter. Only an increase in OmpR phosphorylation induces the expression of CI-LVA that represses the constant and mediated via the λ-phage P_L promoter-expression of the GFP reporter gene (Figure 3.A.). Additionally, the protease-recognized LVA peptide, which is an 11 amino acid long tail fused to the CI protein, increases the sensitivity of the system to changes by reducing the half-life of the CI protein.

The background of the system was identified by measuring the fluorescence of the JM1012 cells without any addition. The background was finally used as a baseline to distinguish between the phosphorylation or dephosphorylation of the OmpR response regulator. Even though we tested both strong and weak ribosome binding sites (RBS) upstream the P_L promoter, with the weak RBS performing better.

All constructs were tested in this cell system under stable temperature conditions (37^{o}C). The exponential increase or decrease of the fluorescence observed was visualized logarithmically by directly comparing the signal to the background baseline, set to 1 (Figure 3.B.). The constructs with an phosphatase activity gave a signal above 1, while the ones with an phosphorylase activity gave signals below 1.

Among the controls used, Taz was the only one with a phosphorylation activity. Any alteration of the system, like the removal of the sensor-HAMP domains, or the removal of the ATP binding domain, switched the functionality of the protein mostly towards a phosphatase activity, even when only the DHp domain was tested.

Amazingly, the fused protein DHp-DX1 showed also a phosphorylase activity that approximated 50% of Taz. Even more interesting was that the mutation of DHp-DX1 on the targeted for phosphorylation histidine (DHp-DX1(H15Q)) could completely reverse the functionality from an OmpR phosphorylase to an OmpR phosphatase. DHp-DX2 had also a similar, yet weak phosphorylase

activity, while no significant function was observed for DX-DHp (Supplementary Figure 2.).

Biophysical protein characterization

In vitro all DX proteins were labile. The TRX fusion improved the stability of the purified proteins DHp-DX1 and 2, while it was not efficient in stabilizing DX-DHp. The buffer used was also a significant parameter for keeping the proteins from precipitating during concentration or the cleavage of the TRX fusion. Further testing showed that the induction with 2% ethanol during the expression of DHp-DX1, lead to expression of highly stable protein, probably due to the ethanol induced heat shock protein expression that facilitates protein folding[53].

Size exclusion chromatography and circular dichroism (DC) verified the stability improvement of the TRX fused DHp-DX1, which showed a well-folded dimeric nature with thermostability up to 67^{o}C. The same expression method was used for the stabilization of the histidine deficient DHp-DX1(H15Q) mutant as well as the DHp-DX2 protein, with the same result. However, the ethanol induction failed to improve the stability of the DX-DHp construct. The DHp-DX1 protein generally was more stable after the removal of the TRX fusion. Further analysis of the cleaved protein indicated that it was forming a dimer, as indicated by 1D NMR (data not shown).

In gel functional detection

To investigate the mechanism of the *in vivo* observed OmpR phosphorylase activity of DHp-DX1, purified protein was tested for histidine auto-phosphorylation, while the DHp-CA protein was used as a positive control. The TRX fused proteins were immobilized under native conditions to visualize the histidine phosphorylation using an N1-anti-phospho-His antibody through western blot. Only the positive control DHp-CA showed a detectable phosphorylation on the N1 position of the histidine (Supplementary Figure

3.A.), while DHp-DX1 showed no signal. Autoradiography was also employed for the detection of any phosphorylation on either the DHp-DX1 protein, or the OmpR response regulator. No significant signal was observed, even when the TRX fusion was removed (Data not shown).

The TRX fused proteins DHp-DX1 and DHp-CA, which was used as a negative control, were further tested for a possible ATPase activity via the in gel ATPase assay of lead phosphate[49]. The proteins were immobilized in a gel under native conditions. The gel was incubated in the presence of 4mM ATP (30mM Tris, 1.5mM MgCl2, 1.5mM ZnCl2, pH 8.5) for 60min at RT and under shaking. The gel was washed with 35mM Tris, pH 8.5 and incubated over night in 35mM Tris, 270mM glycine, 20% v/v Methanol, 0.075% w/v $Pb(NO_3)_2$, 0.8mM ATP, 14mM MgSO4, 14mM ZnSO4, pH 8.5, at -4oC and under shaking. The released free phosphate, formed white bands of insoluble lead phosphate in the gel, which were finally visualized as dark brown bands, using $(NH_4)_2S$ (1% v/v).

The gel was then cut into two parts; with both parts having the same samples we used one for a western blot, while the other half was dyed using coomassie. Each step of the treatment was imaged (Figure 4.A). The in gel ATPase assay in combination with the detection of the proteins of interest with anti-His Tag antibodies could verify that there is a release of free phosphate specifically from the DHp-DX1 protein. The experiments were repeated also for the histidine deficient DHp-DX1(H15Q) mutant, showing no significant changes in the ATPase activity (Supplementary Figure 3.B.). The results above strongly indicated that the DHp-DX1 protein has an ATPase activity that is independent from the histidine found on the DHp domain.

ATPase functional characterization

The DX protein was previously hypothesized to have an ATPase activity when ADP was found bound on the protein during crystallization[41,54], however the ATPase function of DX was not demonstrated biochemically. To determine whether the ATPase activity was specific for the DHp-DX1 protein and was not an effect of the DX protein alone, we used the TRX fused DX protein as a

control. ATPase activity was assessed by the colorimetric malachite green assay. The function of both proteins was screened under different pH, temperature and added metal conditions for different time points.

The TRX-DX fusion showed no ATPase activity under any condition. In contrast, the DHp-DX1 protein hat a Mg^{2+} dependent ATPase activity which was optimized at 28^{o}C and pH equal to 6 (Figure 4.B. and C.). This indentifies the ATPase activity as a novel function, product of the fusion of the DHp and DX domain. The enzymatic activity was kinetically characterized using a continuous-spectrophotometric method (MESG/PNP-dependant phosphate assay; Figure 4.D.). The Vmax, Km and Kcat were calculated to be equal to 0.08267μM/min, 152.1μM and 0.041335 min^{-1}, respectively, indicating a reasonable ATPase activity. Experiments using ADP instead of ATP showed no phosphatase activity, indicating a hydrolysis enzyme specificity to the γ-phosphate moiety of the ATP (Supplementary Figure 4). Experiments with GTP were also performed showing a slower yet detectable GTP hydrolysis (Data not shown).

Discussion

The CA domain has an interesting evolutionary history during which it can be involved in proteins that hydrolyze ATP, or in proteins that transfer the γ-phosphate onto a serine or a histidine residue. In histidine kinases type I the CA domain is interacting with the DHp domain providing the ATP that mediates phosphorylation of the histidine in DHp, which will subsequently lead to phosphorylation of an aspartate in the response regulator. Here we show that an artificial protein, like DX, can replace the CA domain in providing the ATP molecule to the system. The novel fusion protein DHp-DX1 has *in vitro* ATPase activity. This ATPase activity simulates the efficiency of a natural ATPase showing a pH, temperature and Mg^{2+} preference. This activity was absent in the TRX-DX fusion, showing that DX has no ATPase activity independently from the DHp domain. The hydrolysis is specific for the γ-phosphate of the ATP since no phosphate release was observed when ADP was used. The hydrolysis did not involve the histidine residue on DHp.

The ATPase activity observed is common between our fusion protein DHp-DX1 and some CA-containing proteins[33,55]. Furthermore, CA-containing kinases have a similar function were the phosphate is instead transferred onto a serine[32,34] or a histidine residue[44], like in the case of EnvZ. This indicates that probably the ATPase activity was the first step in a functional evolution during the fusion of CA with other domains, serving as a functional stepping stone to the evolution of the serine and histidine kinases. The fact that an analogous ATP binding domain could replace CA indicates that the initial selection of this domain was probably a matter of luck, availability and opportunity.

It is important to point out that even though DX and CA have no evolutionary or sequence similarities, DX shows some structural similarities to the CA domain. The 3 β-sheets and the α1 helix of the DX protein form a similar to the CA domain fold, which includes the participating in the ATP binding site β4, β5 and β7 sheets, as well as the α2 helix (Figure 1). This mind-boggling structural similarity was detected using the online-structural comparison platform DALI[56], where the DX protein structure (Protein data bank or PDB ID:

3ltc) was compared with the whole PDB database. DX showed similarities for multiple CA related domains, like the CA domain found in the blue-light-activated histidine kinase 2 (4r3a-A; Supplementary Figure 5.A.) or even similar to the CA domains in serine/threonine-protein kinase BRSK1 and 2 (5iri-A, 4yom-A; Supplementary Figure 5.C.).

However, this structural similarity includes the helix α1 of the DX protein instead of the helix α2, which participates in the ATP binding site. This is because the distance between the β-sheets and α1 helix better fits the distance between the analogous β-sheets and α2 helix of the CA protein. The DX α2 helix neighboring the ATP binding site is actually more distant from the β-sheets, which along with a short, in comparison to the one on the CA domain, loop fails to properly seal the ATP ligand leaving it unusually accessible to water. This positioning of the ATP probably facilitates the ATP hydrolysis observed for DHp-DX1 and it might be the main factor that interferes with a possible phosphorylation of the histidine residue on the DHp domain.

The fact that no phosphorylation was detected *in vitro* for either the DHp-DX1 protein or the OmpR response regulator, indicates that our protein is predominately functioning as an ATPase*. In parallel, o*ur *in vivo* findings show that in the cell there is a histidine dependent triggering of the gene expression through the ompC promoter, which depends on the phosphorylation of the OmpR response regulator. This effect cannot be a result of the ATPase activity of the protein, since it is completely abolished for the histidine deficient DHp-DX1(H15Q) mutant, while the ATPase activity of the mutant is still present. We hypothesize that our dimeric DHp-DX1 protein allows the interaction between the DHp and the DX domains in a way that not only allows the hydrolysis of the ATP *in vitro*, but also promotes phosphorylation of the response regulator OmpR *in vivo*. This kinase activity might target directly the Asp residue of OmpR, or the His residue on the DHp domain, which seems to play a significant role in the possible phosphotransfer mechanism.

Additionally, even though the Taz chimeric kinase acts predominantly as an OmpR phosphorylase, the independent transmitter DHp-CA has shown a predominant phosphatase activity under the same conditions, *in vivo*. Our

synthetic analog DHp-DX1, on the other hand, demonstrates a possible OmpR phosphorylase activity without the need of a sensor protein.

The ability to detect an effect of possible OmpR phosphorylation *in vivo* has offered the opportunity to further evolve our novel enzyme using *in vivo* evolution by coupling the system to a gene essential for life, mimicking the way natural selection enforces the evolution of highly efficient enzymes. Alternatively, selected mutations could be introduced through design in order to further seal the ATP binding pocket by reducing the distance between the α2 helix and the β-sheets of the DX domain. Another approach could be the refining of the interface between the DHp and DX domains. In addition a longer loop could be introduced to simulate the long loop that functions as a highly flexible lid for the CA ATP binding pocket.

In conclusion, this project has offered a proof of concept of the evolution of primordial enzymes through simulating the evolution of histidine kinases starting with a randomly evolved ATP binding protein. Our protein is a bifunctional ATPase-OmpR phosphorylase precursor of histidine kinases that can be used to further evolve a more specific and efficient kinase activity through *in vivo* evolution or protein design. Finally our approach could be used to develop more novel enzymes by a combination of natural domain recycling, *de novo* evolution and further *in vivo* selection, significantly assisting novel enzyme design.

Contributions

AL conceived the project, IK designed and executed the experiments, ES provided input and ideas regarding the in vivo experiments, HZ provided input and ideas regarding the *in silico* protein design, MC did the NMR analysis. The manuscript and the figures were prepared by IK and critically examined and edited by AL and ES. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Prof. Dr. Volkmar Braun and Prof. Dr. Joachim E. Schultz for offering critical thinking and suggestions during the progress of the project. We are grateful to Dr. Harshul Arora Verasztó and Maria Logotheti for their feedback during the kinetic analysis. We also thank Dr. Tjeerd Dijkstra and Dr. Andre Noll for reading and editing the manuscript. We would finally like to thank Prof. Sung Kuk Lee from the Ulsan National Institute of Science and Technology (UNIST, Korea) for providing us with the JM1012, and pJM1,2 constructs.

References

1. Alberts, B. *et al.* The RNA World and the Origins of Life. *Mol. Biol. Cell 4th Ed.* (2002).
2. Robertson, M. P. & Joyce, G. F. The Origins of the RNA World. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).
3. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28**, 411–418 (2003).
4. Zhang, B. & Cech, T. R. Peptide bond formation by in vitro selected ribozymes. *Nature* **390**, 96–100 (1997).
5. Noller, H. F., Hoffarth, V. & Zimniak, L. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256**, 1416–1419 (1992).
6. Szathmáry, E. The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends Genet. TIG* **15**, 223–229 (1999).
7. Wilson, T. J. & Lilley, D. M. J. RNA catalysis—is that it? *RNA* **21**, 534–537 (2015).
8. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
9. Szilágyi, A., Kun, Á. & Szathmáry, E. Early evolution of efficient enzymes and genome organization. *Biol. Direct* **7**, 38 (2012).
10. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
11. Kim, H. S., Mittenthal, J. E. & Caetano-Anollés, G. MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics* **7**, 351 (2006).
12. Lazcano, A. & Miller, S. L. On the Origin of Metabolic Pathways. *J. Mol. Evol.* **49**, 424–431 (1999).
13. Caetano-Anollés, G., Kim, H. S. & Mittenthal, J. E. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci.* **104**, 9358–9363 (2007).
14. Newton, M. S., Arcus, V. L., Gerth, M. L. & Patrick, W. M. Enzyme evolution: innovation is easy, optimization is complicated. *Curr. Opin. Struct. Biol.* **48**, 110–116 (2018).
15. Tyzack, J. D., Furnham, N., Sillitoe, I., Orengo, C. M. & Thornton, J. M. Understanding enzyme function evolution from a computational perspective. *Curr. Opin. Struct. Biol.* **47**, 131–139 (2017).
16. Copley, S. D. Toward a Systems Biology Perspective on Enzyme Evolution. *J. Biol. Chem.* **287**, 3–10 (2012).
17. Wang, M. & Caetano-Anollés, G. The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World. *Structure* **17**, 66–78 (2009).
18. Wang, M., Yafremava, L. S., Caetano-Anollés, D., Mittenthal, J. E. & Caetano-Anollés, G. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17**, 1572–1585 (2007).
19. Capra, E. J. & Laub, M. T. Evolution of two-component signal transduction systems. *Annu. Rev. Microbiol.* **66**, 325–347 (2012).
20. Dutta, R., Qin, L. & Inouye, M. Histidine kinases: diversity of domain organization. *Mol. Microbiol.* **34**, 633–640 (1999).
21. Zhu, Y., Qin, L., Yoshida, T. & Inouye, M. Phosphatase activity of histidine kinase EnvZ without kinase catalytic domain. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 7808–7813 (2000).
22. Cock, P. J. A. & Whitworth, D. E. Evolution of Prokaryotic Two-Component System Signaling Pathways: Gene Fusions and Fissions. *Mol. Biol. Evol.* **24**, 2355–2357 (2007).
23. Whitworth, D. E. & Cock, P. J. A. Evolution of prokaryotic two-component systems: insights from comparative genomics. *Amino Acids* **37**, 459–466 (2009).

24. Hu, X., Wang, H., Ke, H. & Kuhlman, B. High-resolution design of a protein loop. *Proc. Natl. Acad. Sci.* **104**, 17668–17673 (2007).
25. Lo, J., Tessa Van Tol, Stephanie Yeung & Kevin Zou. EnvZ is not essential for the upregulation of OmpC following treatment with sublethal kanamycin in Escherichia coli. *J. Exp. Microbiol. Immunol. J EMI* **18**, 5 (2014).
26. Smarajit Chakraborty, R. S. W., Leslie K. Morgan, J. Y. & Linda J. Kenney. Non-canonical activation of OmpR drives acid and osmotic stress responses in single bacterial cells.
27. Utsumi, R. *et al.* Activation of bacterial porin gene expression by a chimeric signal transducer in response to aspartate. *Science* **245**, 1246–1249 (1989).
28. Levskaya, A. *et al.* Synthetic biology: Engineering *Escherichia coli* to see light. *Nature* **438**, 441–442 (2005).
29. Michalodimitrakis, K. M., Sourjik, V. & Serrano, L. Plasticity in amino acid sensing of the chimeric receptor Taz. *Mol. Microbiol.* **58**, 257–266 (2005).
30. Lee, J. M., Lee, J., Kim, T. & Lee, S. K. Switchable gene expression in Escherichia coli using a miniaturized photobioreactor. *PloS One* **8**, e52382 (2013).
31. Ferris, H. U. *et al.* The Mechanisms of HAMP-Mediated Signaling in Transmembrane Receptors. *Structure* **19**, 378–385 (2011).
32. Harris, R. A. *et al.* A new family of protein kinases--the mitochondrial protein kinases. *Adv. Enzyme Regul.* **35**, 147–162 (1995).
33. Prodromou, C. *et al.* Identification and Structural Characterization of the ATP/ADP-Binding Site in the Hsp90 Molecular Chaperone. *Cell* **90**, 65–75 (1997).
34. Machius, M., Chuang, J. L., Wynn, R. M., Tomchick, D. R. & Chuang, D. T. Structure of rat BCKD kinase: nucleotide-induced domain communication in a mitochondrial protein kinase. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 11218–11223 (2001).
35. Campbell, E. A. *et al.* Crystal structure of the Bacillus stearothermophilus anti-sigma factor SpoIIAB with the sporulation sigma factor sigmaF. *Cell* **108**, 795–807 (2002).
36. Min, K. T., Hilditch, C. M., Diederich, B., Errington, J. & Yudkin, M. D. Sigma F, the first compartment-specific transcription factor of B. subtilis, is regulated by an anti-sigma factor that is also a protein kinase. *Cell* **74**, 735–742 (1993).
37. Elich, T. D. & Chory, J. Phytochrome: If It Looks and Smells Like a Histidine Kinase, Is It a Histidine Kinase? *Cell* **91**, 713–716 (1997).
38. Bergerat, A. *et al.* An atypical topoisomerase II from Archaea with implications for meiotic recombination. *Nature* **386**, 414–417 (1997).
39. Mechaly, A. E., Sassoon, N., Betton, J.-M. & Alzari, P. M. Segmental helical motions and dynamical asymmetry modulate histidine kinase autophosphorylation. *PLoS Biol.* **12**, e1001776 (2014).
40. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
41. Simmons, C. R. *et al.* A Synthetic Protein Selected for Ligand Binding Affinity Mediates ATP Hydrolysis. *ACS Chem. Biol.* **4**, 649–658 (2009).
42. Smith, M. D. *et al.* Structural Insights into the Evolution of a Non-Biological Protein: Importance of Surface Residues in Protein Fold Optimization. *PLoS ONE* **2**, e467 (2007).
43. Simmons, C. R. *et al.* Three-Dimensional Structures Reveal Multiple ADP/ATP Binding Modes for a Synthetic Class of Artificial Proteins,. *Biochemistry* **49**, 8689–8699 (2010).
44. Ferris, H. U., Coles, M., Lupas, A. N. & Hartmann, M. D. Crystallographic snapshot of the Escherichia coli EnvZ histidine kinase in an active conformation. *J. Struct. Biol.* **186**, 376–379 (2014).

45. Chaput, J. C. & Szostak, J. W. Evolutionary Optimization of a Nonbiological ATP Binding Protein for Improved Folding Stability. *Chem. Biol.* **11**, 865–874 (2004).
46. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607–618 (2011).
47. Andersen, J. B. *et al.* New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl. Environ. Microbiol.* **64**, 2240–2246 (1998).
48. Ibrahim, I. M., Puthiyaveetil, S. & Allen, J. F. A Two-Component Regulatory System in Transcriptional Control of Photosystem Stoichiometry: Redox-Dependent and Sodium Ion-Dependent Phosphoryl Transfer from Cyanobacterial Histidine Kinase Hik2 to Response Regulators Rre1 and RppA. *Plant Cell Biol.* 137 (2016). doi:10.3389/fpls.2016.00137
49. Suhai, T., Heidrich, N. G., Dencher, N. A. & Seelert, H. Highly sensitive detection of ATPase activity in native gels. *Electrophoresis* **30**, 3622–3625 (2009).
50. Ferris, H. U. *et al.* Mechanism of Regulation of Receptor Histidine Kinases. *Structure* **20**, 56–66 (2012).
51. Igo, M. M., Ninfa, A. J., Stock, J. B. & Silhavy, T. J. Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor. *Genes Dev.* **3**, 1725–1734 (1989).
52. Sourjik, V. & Berg, H. C. Receptor sensitivity in bacterial chemotaxis. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 123–127 (2002).
53. Georgiou, G. & Valax, P. Expression of correctly folded proteins in Escherichia coli. *Curr. Opin. Biotechnol.* **7**, 190–197 (1996).
54. Surdo, P. L., Walsh, M. A. & Sollazzo, M. A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nat. Struct. Mol. Biol.* **11**, 382–383 (2004).
55. Dutta, R. & Inouye, M. GHKL, an emergent ATPase/kinase superfamily. *Trends Biochem. Sci.* **25**, 24–28 (2000).
56. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44**, W351–W355 (2016).

Figures

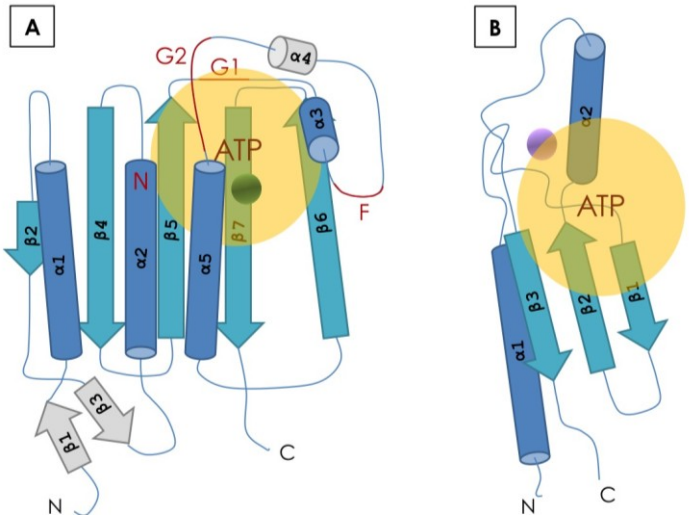


Figure 1. Cartoon representation of the CA and DX protein structures. (A) CA domain structural representation. The gray secondary structure elements are unique to the EnvZ CA domain, the blue ones are conserved structural elements found in the known Bergerat fold. The conserved sequence boxes N, G1, F and G2 are shown in red color. The flexible long loop between F and G2 boxes acts as a lid sealing the ATP binding pocket. **(B)** DX domain structural representation. The protein forms a zinc finger. The ATP is quite exposed and bound between α-helix α2 and the β-sheets β1-β3. No Mg^{2+} was found bound to the DX structure. The approximate position of the ATP binding pocket is presented as orange shadow. The Mg^{2+} and Zn^{2+} metals are represented as green and purple spheres, respectively.

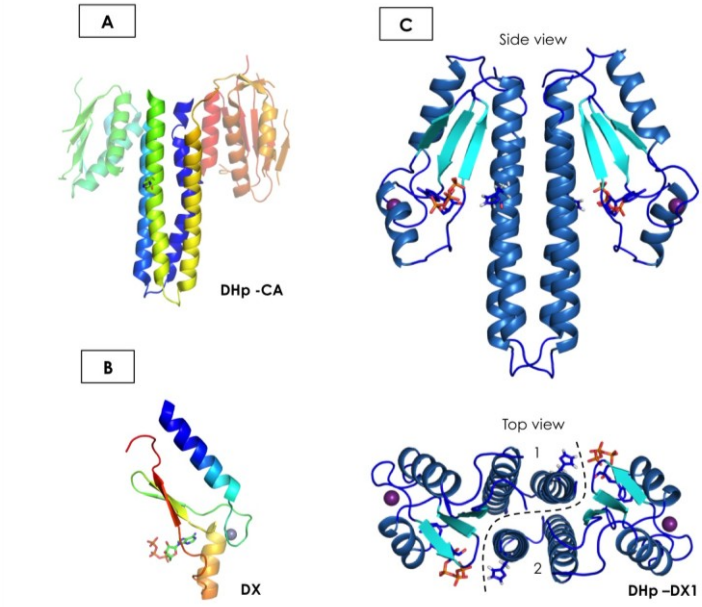


Figure 2. The protein model of DHp-DX1. (A) Structure of the dimeric DHp-CA transmitter of the EnvZ histidine kinase in *E. coli* (The DHp domain is highlighted). **(B)** Structure of the *In virto* evolved DX protein. The rainbow color from blue to red indicates the N to C-terminal orientation of the proteins. **(C)** Structural model of the dimeric DHp-DX1 protein, front and top view. DHp-DX1 was the most promising fusion protein of the DHp and DX domains. The fusion included the natural EnvZ linker connecting the two domains, which were fused in the natural domain order. Loops, helices and sheets are shown in different shades of blue.

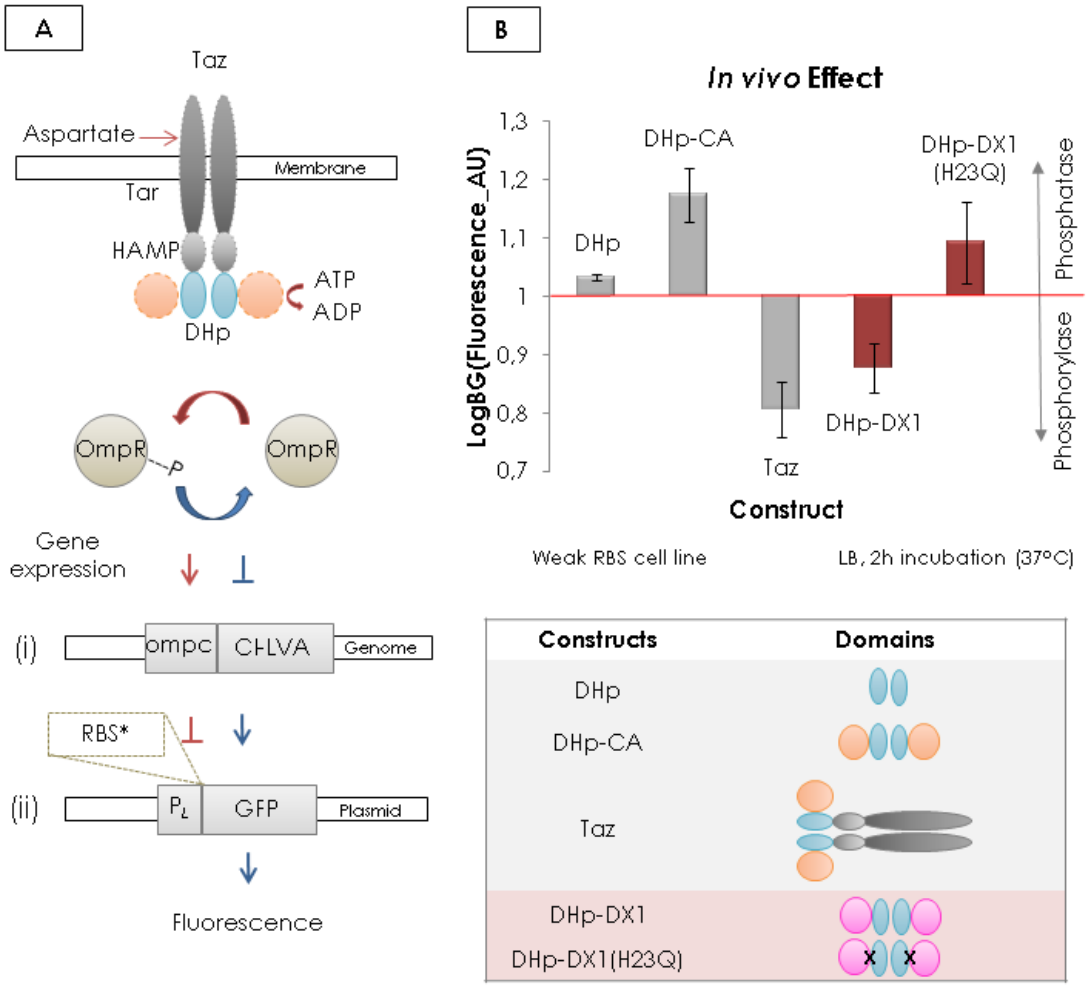


Figure 3. ***In vivo*** **functional analysis of the DHp-DX1 protein. (A)** Adaptation from the Taz and Cph1-chimeric TCST systems, allowing the detection of both OmpR phosphorylation and dephosphorylation based on the dual control of the expression of the GFP reporter gene. In the first step the ompC promoter controls the expression of the CI_LVA repressor (i) which then represses the always activated and mediated via the P_L promoter GFP expression. Domain removals or exchanges are shown with dotted outlines. Both weak and strong RBS available(*). **(B)** Graphic representation of the GFP fluorescence as a logarithm, using the fluorescence background as a baseline. The cell line with the weak RBS gave the best results. The assay conditions are shown on the bottom of the graph. OmpR phosphatase activity increases the fluorescence (upward bars) while OmpR phosphorylase activity decreases the fluorescence (downward bars). The constructs shown are listed in the table, grey represents controls used while red indicates the fusion DHp-DX1 and the DHp-DX1(H15Q) histidine deficient mutant. DHp-DX1 showed histidine dependant OmpR phosphorylase activity *in vivo.* The experiment was repeated three times including 6 sample replications.

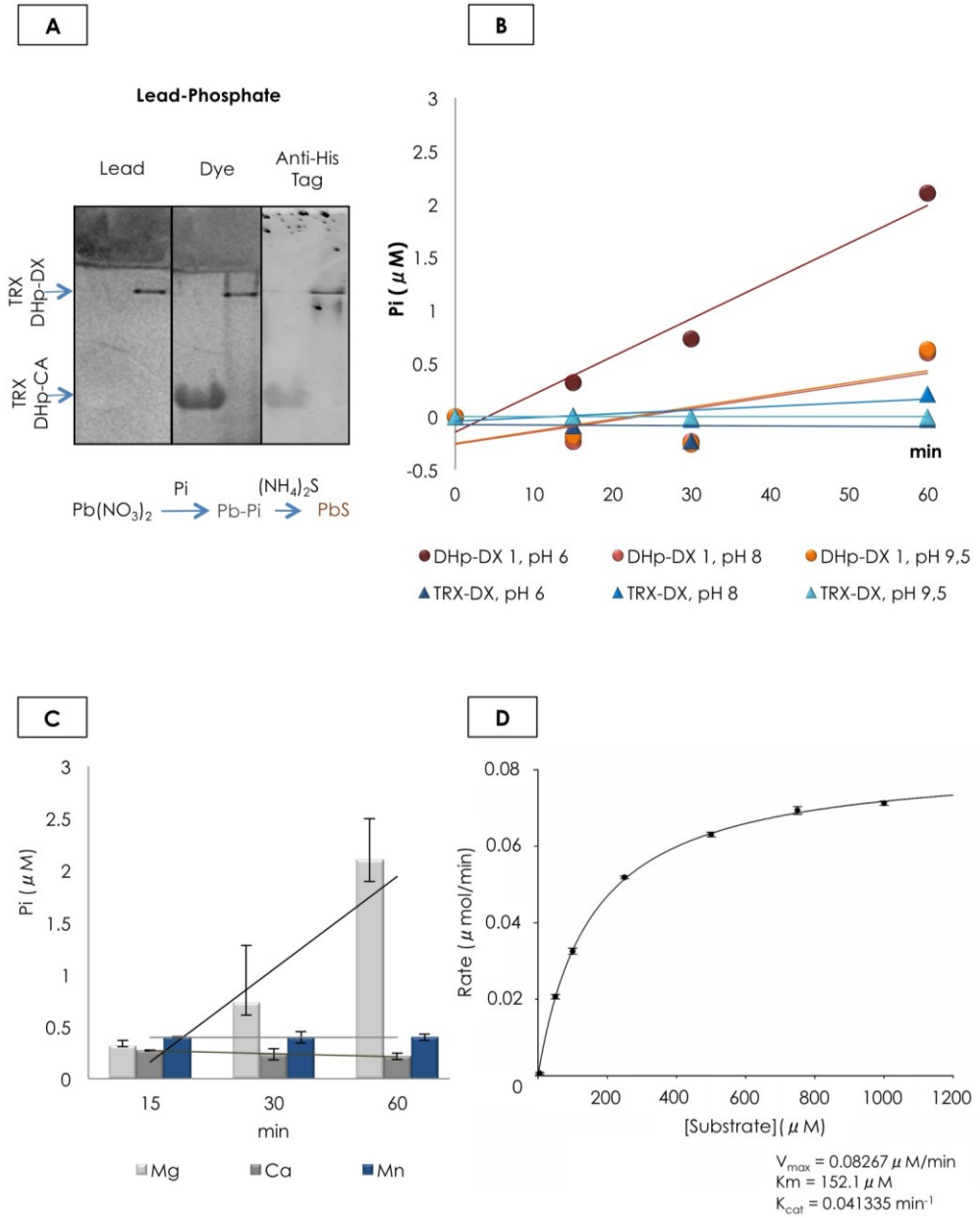


Figure 4. ***In vitro*** **ATPase functional characterization of DHp-DX1. (A)** The TRX fused DHp-CA and DHp-DX1 proteins were immobilized in gel, under native conditions. The gel was imposed to the lead-phosphate reaction overnight, leading to the formation of a brown band were free phosphate was released, and later dyed using coomassie. DHp-DX1 showed an ATPase activity, while the DHp-CA natural analog did not. The same samples were used for western blot with anti-his tag antibodies to verify that the signal comes from the proteins of interest. **(B)** The ATPase activity was further tested under different pH conditions (6, 8 and 9.5) with the malachite green reaction at room temperature and in the presence of 10mM $MgCl_2$. Both the DHp-DX1 (warm shade circles) and the TRX-DX (cold shade circles) fusion were used. Only the DHp-DX1 protein showed significant ATPase activity, especially under a pH equal to 6. **(C)** The same experiment was repeated with the addition of different metals. All experiments were repeated four times. **(D)** Michaelis-Menten graphical representation of the DHp-DX1 ATPase kinetics. The continuous, spectrophotometric and MESG/PNP(2-amino-6-mercapto-7-methylpurine riboside/ purine nucleoside phosphorylase)-dependant phosphate assay was used for the ATPase kinetic analysis of the DHp-DX1 protein. The test was performed under optimal conditions (2μM protein in pH 6, 28°C in the presence of 10mM $MgCl_2$). The catalytic activity of the enzyme was quite high giving an Vmax, Km and Kcat estimation of 0.08267μM/min, 152.1μM and 0.041335 min^{-1}, respectively.

Supplementary material

Sequences

>DHp-DX1_pcl1920_plasmid

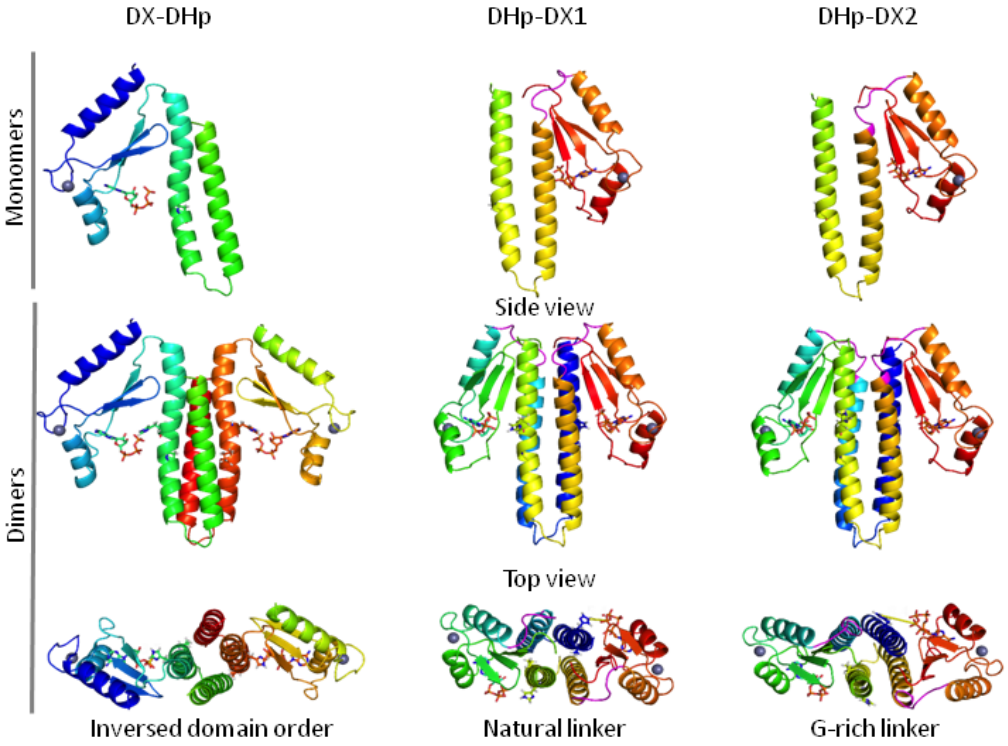
ATGATTACGCCAAGCTTGCAGTTAGCCGATGATCGCACCCTGTTAATGGCTGGTGTGAGTCATGATCTC
CGTACACCGCTCACCCGGATTCGCCTGGCAACCGAAATGATGAGCGAACAAGACGGCTATCTGGCAGAA
TCGATTAACAAAGATATTGAGGAGTGTAATGCCATCATCGAACAGTTTATCGACTATCTGCGTACTGGC
CAGGAAATGCCGGGTGACAAGAAAACGAATTGGCTGAAACGCATTTACCGTGTACGCCCATGTGTGAAA
TGCAAAGTTGCACCTCGTGACTGGAAAGTCAAGAACAAACACCTTCGGATTTACAACATGTGCAAAACG
TGCTTTAACAACTCCATTGACATTGGTGACGATACCTATCATGGCCATGTGGATTGGCTGATGTATGCG
GATTCATAA

>DHp-DX1_pCL1920_protein

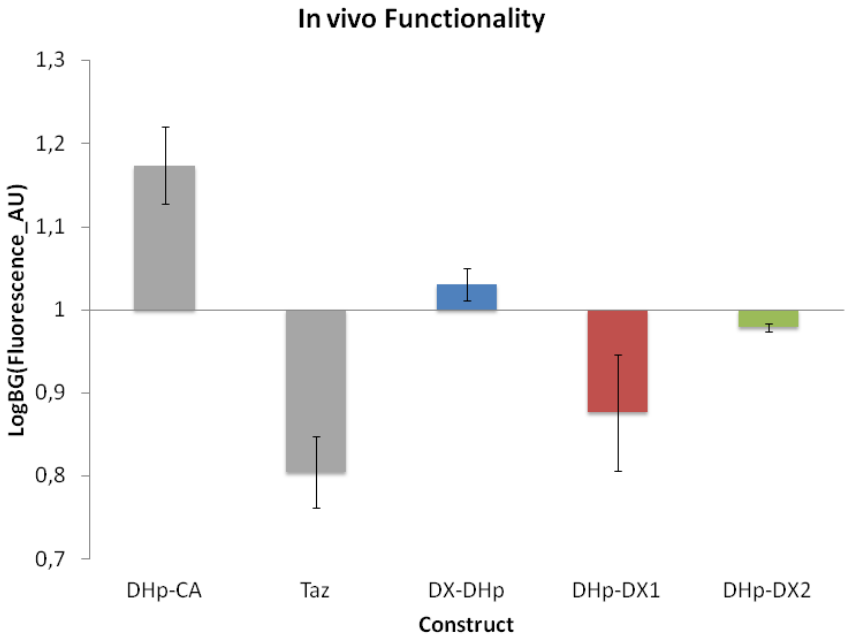
MTMITPSLQLADDRTLLMAGVSHDLRTPLTRIRLATEMMSEQDGYLAESINKDIEECNAIIEQFIDYLR
TGQEMPGDKKTNWLKRIYRVRPCVKCKVAPRDWKVKNKHLRIYNMCKTCFNNSIDIGDDTYHGHVDWLM
YADS

Supplementary Table 1. Primers used in the present study

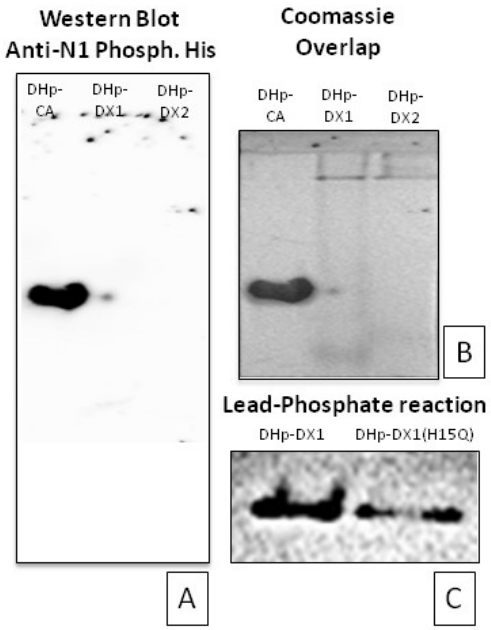
CA_R_BamHI	AGTGAATTCGAGCTCGGTA
DHP_R_BamHI	TTATGGATCCTTAGCGCAGGTAGTCGATAAACTGCT
DHp-DX1_Q_F	ATGGCTGGTGTGAGTCAaGATCTCCGTACACCGCT
DHp-DX1_Q_R	AGCGGTGTACGGAGATCtTGACTCACACCAGCCAT
DHp-DX1_R_BamHI	TTATCCGGATCCTTATGAATCCGCATACATCAGC
DHp-DX2_R_BamHI	TTATCCGGATCCTTAACTATCGGCGTACATCAGC
DX_R_BamHI	TTATCCGGATCCTTAATCCGCATACATCAGCCAGTC
DX-DHp_R_BamHI	TTATCCGGATCCTTAGCGCAGATAGTCGATAAACTGCTCAATGA
pCL1920/DHp_F_HindIII	TTATCCAAGCTTGGCGGCTGGTGTTAAGC
pCL1920/DHp-DX1_F_HindIII	TTATCCAAGCTTGCAGTTAGCCGATGATCGC
pCL1920/DHp-DX2_F_HindIII	TTATCCAAGCTTGCAACTGGCGGATGATCG
pCL1920/DX-DHp_F_HindIII	TTATCCAAGCTTGGATGATGACGACAAGAAAACCAATTGG
pCL1920/Taz_F_HindIII	ACCATGATTACGCCAAGCTT
pETTRX/DHp-CA_F_NcoI	TTATCCATGGGGCAACTGGCGGATGACC
PETTRX/DHp-DX1_F_Ncol	TTATACCATGGGACAGTTAGCCGATGATCGC
PETTRX/DHp-DX2_F_Ncol	TTATACCATGGGACAACTGGCGGATGATCG
PETTRX/DX-DHp_F_Ncol	TTATACCATGGATGATGACGACAAGAAAACCA



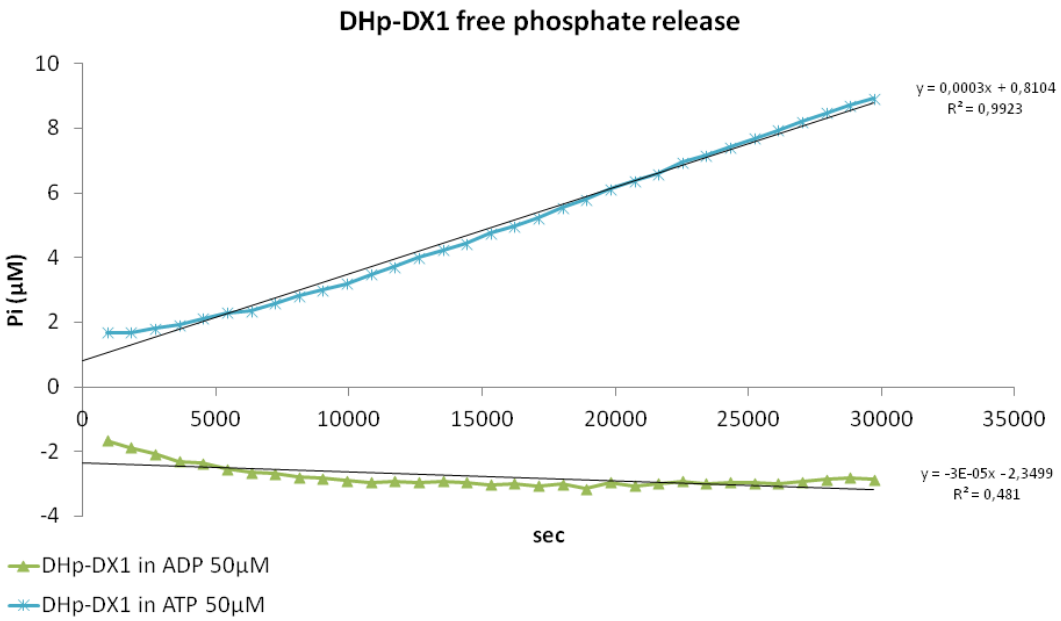
Supplementary Figure 1. **DHp and DX fusion construct design.** Three models were designed. From top to bottom the monomeric and dimeric forms (front and top view) of the proteins are shown. **(A)** DX-DHp protein with an reversed domain orientation (cis-autophosphorylation). **(B)** DHp-DX1 protein with a DHp-natural_linker-DX natural like domain orientation (trans-autophosphorylation). **(C)** DHp-DX2 protein with a DHp-G/rich_linker-DX natural domain orientation (trans-autophosphorylation). The rainbow color from blue to red indicates the N to C-terminal orientation of the proteins. The loops are shown in magenta.



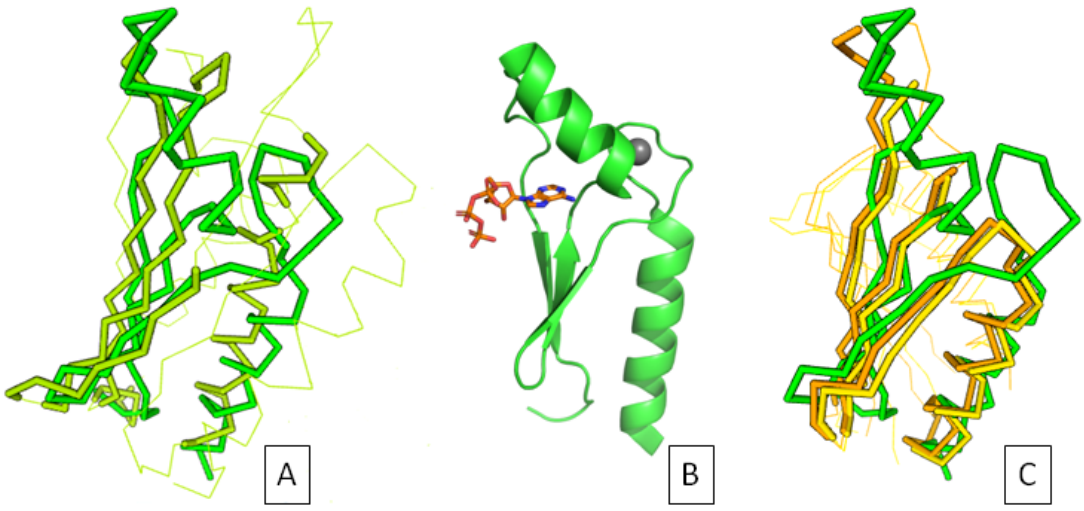
Supplementary Figure 2. ***In vivo*** **functional analysis of the DHp, DX fused proteins.** Graphic representation of the GFP fluorescence as a logarithm, using the fluorescence background as a baseline. OmpR phosphatase activity increases the fluorescence (upward bars) while OmpR phosphorylase activity decreases the fluorescence (downward bars). The constructs shown in grey represents controls used while blue, red and green indicate the fusion proteins DX-DHp, DHp-DX1 and DHp-DX2, respectively. Only the DHp-DX1 showed a significant OmpR phosphorylase activity. The *in vivo* system used was an adaptation from the Taz and Cph1-chimeric TCST systems, allowing the detection of both OmpR phosphorylation and dephosphorylation. The experiment was repeated three times including 6 sample replications.



Supplementary Figure 3. **In gel functional analysis. (A)** Western blot with N1 anti-phospho-histidine antibody for the TRX fused proteins DHp-CA, DHp-DX1 and DHp-DX2. Only the positive control shows a detectable histidine phosphorylation on the N1 position. **(B)** Overlapping of the western blot signal with the coomassie dyed gel. **(C)** The Lead phosphate reaction of the TRX fused proteins DHp-DX1 and the histidine deficient DHp-DX1(H15Q). The ATPase activity does not depend on the histidine of the DHp domain.



Supplementary Figure 4. Comparison of the DHp-DX1 hydrolytic activity in the presence of either ATP or ADP. Graphical representation of the DHp-DX1 free phosphate release using the continuous, spectrophotometric and MESG/PNP(2-amino-6-mercapto-7-methylpurine riboside/ purine nucleoside phosphorylase)-dependant phosphate assay. The test was performed under optimal conditions (2µM protein in pH 6, 28°C in the presence of 10mM $MgCl_2$) with the addition of either ATP of ADP (50µM).



Supplementary Figure 5. Structural similarities between DX and CA related domains. **(A)** Superimposition of DX and the CA domain of the blue-light-activated histidine kinase 2 (4r3a-A, dark green) **(B)** Cartoon representation of the DX protein, provided for comparison. **(C)** Superimposition of DX with similar to the CA domains in serine/threonine-protein kinase BRSK1 and 2 (5iri-A, orange; 4yom-A, yellow). The DX protein is always presented in the same light green color. The superimposition was accomplished using the online structural comparison platform Dali[56].